THE PHYSICAL SOCIETY

REPORTS ON PROGRESS IN PHYSICS

VOLUME XVI (1953)

EXECUTIVE EDITOR
A. C. STICKLAND

Price to non-Fellows £3 3s. net (Postage 2s.)
Price to Fellows (one copy) £1 7s. 6d. net (Postage 2s.)

Published by
THE PHYSICAL SOCIETY
1 Lowther Gardens, Prince Consort Road
London S.W. 7
1953

REPRINTED BY LITHOGRAPHY IN GREAT BRITAIN
BY JARROLD AND SONS, LTD, NORWICH

CONTENTS

	PAGE
Preface	iv
Neutron Diffraction. By G. E. Bacon and K. Lonsdale	1
Physical Properties and Atomic Arrangements in Crystals. By W. A. Wooster	62
Raman Effect in Solids. By A. C. Menzies	83
Paramagnetic Resonance. By B. Bleaney and K. W. H. Stevens	108
Semiconductor Circuit Elements. By J. S. Blakemore, A. E. De Barr and J. B. Gunn	160
Electrical Discharges. By F. Llewellyn Jones	216
Fluctuation Theory in Physical Measurements. By C. W. McCombie	266
Cosmology. By W. H. McCrea	321
The New Unstable Cosmic-Ray Particles. By G. D. Rochester and C. C. Butler	364

PREFACE

The *Reports on Progress in Physics* is again under the general supervision of the Papers Committee of the Physical Society for the selection of authors and subject matter for each volume of the Reports, and the executive editor of the publication is Dr. A. C. Stickland, Secretary–Editor of the Physical Society.

The members of the Papers Committee were as follows:

Chairman:—Professor S. Devons.

Honorary Papers Secretary:—Dr. H. H. Hopkins.

Ordinary Members: — The President (Professor R. Whiddington), Mr. C. H. Collie, Dr. W. Ehrenberg, Mr. D. W. Fry, Dr. E. P. George, Mr. R. W. G. Hunt, Professor F. L. Jones, Dr. G. O. Jones, Professor H. Jones, Dr. H. Kuhn, Professor H. S. W. Massey, Professor R. E. Peierls, Mr. W. West, Dr. H. Wilman and Mr. C. G. Wynne.

A. C. STICKLAND,
Secretary–Editor.

NEUTRON DIFFRACTION

By G. E. BACON* and K. LONSDALE†

* Atomic Energy Research Establishment, Harwell, Didcot, Berks.
† University College, London

CONTENTS

		PAGE
§1.	Introduction	2
	1.1. Monochromatic neutron beams	3
	1.2. Apparatus and techniques	3
§2.	Neutron scattering by atoms and microcrystals	4
	2.1. Nuclear scattering	5
	2.2. Scattering by microcrystals	11
	2.3. Magnetic scattering	14
	2.4. Polarization effects in magnetized samples	18
§3.	Diffraction by crystals	20
	3.1. Experimental measurements with single crystals	25
	3.2. The use of powder methods	26
§4.	Fundamental experimental measurements	29
	4.1. General determinations of the sign of the coherent scattering amplitude $\bar{a}$	29
	4.2. General determinations of the magnitude of $\bar{a}$	30
	4.3. Other methods of determination of sign and magnitude of $\bar{a}$	33
	4.4. Coefficients of true absorption	37
§5.	The role of neutron diffraction in crystallographic analysis and in a study of the physics of the solid state	37
§6.	Crystal structure determinations involving light elements	40
	6.1. Ice	40
	6.2. Hydrides of heavy elements	41
	6.3. Heavy element carbides	42
	6.4. KHF_2, potassium bifluoride	42
	6.5. Ammonium halides	44
	6.6. VN, VC and KN_3	47
§7.	Structural investigations involving distinction between atoms of closely similar atomic number	48
	7.1. Order–disorder phenomena in alloys	48
	7.2. Spinel structures	49
§8.	Neutron diffraction in magnetic materials	49
	8.1. Paramagnetic materials	50
	8.2. Antiferromagnetic materials having spin moments only	50
	8.3. Antiferromagnetic materials having spin and orbital moments	52
	8.4. Ferro- and ferrimagnetic materials	53
	8.5. Transmission of neutrons through magnetic materials	55
§9.	Other applications	55
	9.1. Nuclear positions	55
	9.2. Diffraction by liquids	56
	9.3. Diffraction by gases	56
	Acknowledgments	56
	List of symbols	57
	References	58

ABSTRACT. A description is given of the progress over the past few years of neutron diffraction, with particular reference to its applications, often alongside complementary X-ray studies, to investigations in structural crystallography and the physics of the solid state. This outlines the relevant principles of neutron scattering by nuclei and magnetic atoms and the way in which these processes determine the intensities of diffracted beams from single crystals and polycrystalline material. Details are given of the experimental determinations of the fundamental scattering amplitudes and cross sections used in diffraction work, and tables of these quantities are included. A review of the physical and chemical problems on which significant information has already been obtained by neutron diffraction groups these problems under three main headings: structural investigations involving light elements, distinctions between atoms of closely similar atomic number, and studies of magnetic materials with particular reference to ferro-, antiferro- and ferri-magnetism. A list of about 150 references is appended.

§1. INTRODUCTION

THE past few years, with the increasing operation of atomic piles, have seen a rapidly growing interest in the study of the interaction of neutrons with matter. Such studies are of value from two main aspects, first as a source of information about atomic nuclei and about neutrons, and secondly as a means of investigating the way in which atoms build up the different states of matter. It is this second application, largely the use of neutron beams for examining the structure of solids, with which we shall be concerned in the present report. In considering the technique and application of 'neutron diffraction' we shall examine particularly the way in which it has been used to supplement studies by X-ray diffraction and to open up new fields of investigation which are outside the possibilities of the latter.

A number of properties of the neutron make it peculiarly valuable for the study of solids. Being an uncharged particle, a little heavier than a hydrogen nucleus, and of very small dimensions, only about 10^{-4} × the size of the atoms in a crystal, it has a very high penetrating power. Moreover the behaviour of neutrons can be described by wave mechanics, and a wavelength λ can be ascribed, equal to $\boldsymbol{h}/mv$ where $\boldsymbol{h}$ is Planck's constant and m, v are the mass and velocity of the neutron. We shall see that it is relatively easy to produce beams of neutrons of uniform velocity in equilibrium with matter at temperatures T a little above room temperature. For such neutrons $\frac{1}{2}mv^2=\boldsymbol{k}T$, where $\boldsymbol{k}$ is Boltzmann's constant, and it turns out that these neutrons will have a wavelength between 1 and 2 Å. This is about the same as the distance apart of atoms in crystals, and much the same as the wavelengths of the characteristic X-rays used for diffraction studies. On the other hand, whereas the vibration frequency of these X-rays is about 10^{18} $\sec^{-1}$, that of the neutrons is only 10^{13} $\sec^{-1}$, a value which is of some interest since the frequencies of thermal vibrations in solids are of this order.

It was first demonstrated by Halban and Preiswerk (1936) and by Mitchell and Powers (1936) that neutrons could indeed be diffracted by solids, following a theoretical suggestion by Elsasser (1936) that this should be possible. These early experiments were done with radium–beryllium neutron sources, and it is only in later years, with the development of chain-reacting atomic piles, that neutrons have become available in sufficiently intense beams for practical use to be made of diffraction techniques.

1.1. *Monochromatic Neutron Beams*

An essential component of the atomic pile is the moderator of graphite or heavy water which slows down the high velocity neutrons, produced by fission, to roughly thermal velocity, at which they are efficient at producing further fission. The pile is thus a source of thermal neutrons, having velocities of the order of a mile a second, and these can be extracted in the form of a defined beam by inserting a suitable collimator. We shall find that in order to obtain an adequate number of neutrons for powder diffraction the collimator aperture has to be of the order of a few centimetres square, and it must be correspondingly long in order to ensure an adequate resolution in the diffraction patterns. The beam emerging from the collimator will not be monochromatic but will have roughly a Maxwellian energy distribution centred about a wavelength of approximately 1·5 Å. A narrow band of wavelengths can be separated from it, either by using a mechanical velocity selector or, more conveniently for a wavelength of about 1 Å, by a crystal monochromator. The width of this wavelength *band*, which must be contrasted with the characteristic *lines* used in x-ray work, will largely be determined by the collimation. In practice a compromise has to be made between reducing the bandwidth by narrowing the angle of the collimator and, at the same time, maintaining an adequate number of neutrons in the incident beam. Generally a collimation of about $\pm\frac{1}{2}°$ is used, and in the case of a graphite pile, where the total neutron flux before the collimator is of the order of 5×10^{11} neutrons/cm²/sec, the number of neutrons in the beam reflected by the monochromator will be a few million per minute.

1.2. *Apparatus and Techniques*

In the earliest experimental work (Sturm 1947, Wollan and Shull 1948a) metallic halide crystals, particularly LiF, NaCl and CaF_2, were used as monochromators. Sturm (1947) has given comparative intensity data for a number of such crystals. These crystals have mosaic spreads of only a minute or two of arc, and it was shown later by Shull and Wollan (1951) that crystals of metals such as lead and copper gave intensities which were greater by a factor of two or three times, and these are now generally used. The monochromator may be used in either the 'reflection' or 'transmission' position. The latter requires a much smaller crystal for intercepting a given width of beam, but the former is more often used in order to take advantage of the 'foreshortening' technique first demonstrated by Fankuchen (1937) for x-rays. In this case the crystal is cut with the surface inclined at a few degrees to the reflecting planes. In the course of investigations of the properties of monochromators Weiss, Hastings and Corliss (1951) have shown that many large crystals grown from the melt have a lineage structure. The various fibrous regions, which may have orientations differing by about $\frac{1}{2}°$, are themselves built up of mosaic blocks mutually inclined within a few minutes of arc.

Most of the investigations reported in the literature so far have been made with powdered materials, and the diffracted intensities have accordingly been relatively low. A fairly large sample has to be used on this account, usually 2 or 3 cm in height and about 1 cm thick, for a parallel slab arranged in the symmetrical transmission position, or 1 cm or so in diameter for a circular cylinder completely bathed in the beam. Detection of the diffracted neutrons is usually by means of a boron trifluoride proportional counter, of the type

described by Fowler and Tunnicliffe (1950), which is carried on an arm which rotates about the axis of the specimen. In order that the full width of the diffracted beam shall enter the counter the aperture of the latter has to be large. Counters of about 5 cm in diameter are used, and in order to obtain efficiencies of the order of 80% it is necessary for them to be 50 cm in length, even when the boron in the BF_3 is present as the highly absorbing isotope ^{10}B. Angular resolution in the powder spectrum is achieved by setting the aperture of the counter at an appreciable distance, usually about 60 cm from the specimen. This requirement, together with the necessity for reducing the number of extraneous background neutrons entering the counter (which entails surrounding it with several inches of shielding material), results in a very large size for the neutron spectrometer. Massive construction is then necessary to give the required precision of angular movement. Thus, although the principles of the technique are very similar to the corresponding measurements with x-rays, the scale of the apparatus is much larger. Various spectrometers in use in the different laboratories carrying out work on neutron diffraction have been described in the literature (Wollan and Shull 1948 a, Hurst, Pressesky and Tunnicliffe 1950, Bacon, Smith and Whitehead 1950).

Recording of the diffraction pattern is made automatic. The counter arm is generally rotated continuously, at a slow speed, and the diffracted neutrons received by the counter are either recorded at small angular intervals on a printing register, operated by scaling circuits, or indicated on a pen recorder fed by a ratemeter. On account of the low intensities of the diffraction peaks only a very slow speed of rotation, of about five degrees per hour, is permissible.

In addition to the more conventional types of powder spectrometer, Sawyer, Wollan, Bernstein and Peterson (1947) have described an instrument using a bent crystal monochromator, and Chamberlain (1950) has reported an instrument of unusual geometry designed for investigations of liquids.

For single crystal work, which is a field still to be exploited to any appreciable extent, the mechanical requirements of the apparatus are much simplified. This arises since there is no difficulty in getting adequate resolution. Much smaller specimens can be used, indeed we shall see later that this is essential for interpreting the intensity data, and the problem of the overlapping of close lines which occurs for powders does not arise. The only factor which has so far prevented the development of a single crystal neutron diffraction spectrometer of little greater size than a conventional x-ray instrument is that a long counter is necessary for high efficiency. Photographic methods of detection have been tried, such as that using a subsidiary indium foil described by Wollan, Shull and Marney (1948); and a much smaller counter incorporating boron-coated foils has been used by Lowde (1950). These methods present easier shielding problems but do not give the high efficiency of the $^{10}BF_3$ proportional counter. However, current investigations (Bernstein and Schardt 1952, and Schenck and Heath 1952) of the detection of neutrons by activated iodide crystals, such as lithium iodide activated by tin, offer encouraging prospects of the future development of small counters.

§2. NEUTRON SCATTERING BY ATOMS AND MICROCRYSTALS

In seeking to investigate the structure of solids with the help of neutron beams we are called upon to make experimental measurements of the intensity and angular position of diffracted spectra and to deduce from these the atomic arrangement. The angular positions of the spectra will be determined by the

dimensions of the unit cell by which we can describe the three-dimensional periodicity which exists in solids. Their intensities will be determined by the manner in which the atoms are distributed about the unit cell, and the way in which the incident radiation interacts with these individual atoms. It is this latter process, the scattering of radiation by atoms, which is so different for x-rays and neutrons and which is the basis whereby, in certain cases, neutron diffraction may provide structural information which is not available from x-ray studies.

Whereas the process of x-ray scattering by atoms is electronic, the electromagnetic radiation interacting with the cloud of electrons which surrounds the nucleus, neutrons in most cases are scattered only by the nucleus. The important exceptions to this are the magnetic ions which give additional scattering by virtue of interaction of their unbalanced electron spins with the magnetic moment of the neutron. We shall take up the discussion of this effect later in this section. Apart from this, a very small neutron–electron interaction has been deduced from experimental observations and correlated with theoretical prediction (see, for example, Hamermesh, Ringo and Wattenberg 1952, Fried 1952).

2.1. *Nuclear Scattering*

Except in the case of magnetic atoms, we are justified in regarding the neutron scattering as purely nuclear, and this fundamental difference from x-rays immediately leads to several important consequences. First, the atomic scattering factors for x-rays increase progressively throughout the Periodic Classification, in accordance with the continual increase in the number of extra-nuclear electrons. On the other hand, although there is a small increase of nuclear scattering factor for neutrons with mass number, due to increasing nuclear size, nevertheless this is largely hidden by resonance effects which vary in a seemingly arbitrary fashion from atom to atom and, indeed, from isotope to isotope for certain elements. The result of this is that, within a factor of about four times, the neutron scattering factors for different nuclei are in general all of the same order, corresponding roughly in magnitude to the x-ray scattering for an element such as oxygen. Although we remark here, at the outset, on the relative equality of the neutron scattering factors, we shall see later that it is the differences between atom and atom, particularly when these are marked between neighbouring atoms, which lend power to neutron diffraction methods.

The atomic scattering factor for x-rays is markedly angular dependent, falling off according to a 'form factor' with increase of 2θ, the angle between the incident and scattered radiation. This occurs because the cloud of scattering electrons has dimensions of about 1 Å, which is comparable with the x-ray wavelength, with the result that only close to the forward direction will the scattered radiation from various parts of the electron cloud be in phase. The radius of an atomic nucleus is, however, only about 10^{-13} cm, and so is much smaller than the neutron wavelength. Effectively, therefore, the nucleus may be regarded as a point scatterer and there will be no falling off with θ of the neutron scattering factor.

A further dependence on angle of x-ray scattering occurs because of polarization. It can be shown that for a free electron in an electromagnetic field of unit amplitude the scattered amplitude at distance r is equal to

$$\frac{e^2}{mc^2}\frac{1}{r}\sin\chi, \qquad \text{.....(1)}$$

where $-e$, m are the electronic charge and mass, c is the velocity of light and χ is the angle between the direction of travel of the scattered beam and the direction of vibration of the incident radiation. If the direction of vibration lies in the scattering plane then $\sin\chi$ will be equal to $\cos 2\theta$, in contrast to a value of unity when the vibration direction is perpendicular to this plane. Thus in the case of unpolarized X-radiation the intensity is controlled by a polarization factor equal to $\frac{1}{2}(1+\cos^2 2\theta)$. There is no angular-dependent factor of this sort for neutrons.

If for the purpose of neutron scattering the nucleus of an atom could be regarded as an impenetrable sphere, then it would be expected, according to wave mechanical theory (Mott and Massey 1949, p. 38), to offer a cross section for scattering equal to $4\pi R^2$ where R is the nuclear radius. This value is four times as great as would be expected on simple classical theory and applies for neutrons of low velocity such as the thermal neutrons which we are considering. The nuclear radius is approximately equal to $1{\cdot}5\times 10^{-13}A^{1/3}$ cm where A is the mass number, the factor $A^{1/3}$ arising from the fact that all nuclear matter is of constant density. In the absence of any resonance effects, to be explained later, it is found that nuclear scattering cross sections are indeed in fair agreement with the value $4\pi R^2$, which is usually described as the potential scattering cross section. Feshbach, Peaslee and Weisskopf (1947) have given a table illustrating this for the heavier elements. In general, however, and particularly for the lighter elements, the process of nuclear scattering is more complicated than this. It is determined by the boundary conditions at the nuclear surface and is dependent on the nature and position of the energy levels of the unstable compound nucleus formed by combination of the scattering nucleus and the incident neutron. It is these energy levels which determine the probability of the various processes which may occur, such as neutron scattering or absorption, with or without the emission of radiations. To a first approximation the behaviour can be regarded according to the Breit–Wigner (1936) dispersion formula which expresses both scattering and absorption cross sections in terms of the empirical constants of a single resonance energy level. The scattering cross section is given by

$$\sigma=\frac{4\pi}{\kappa^2}\left|\kappa R+\frac{\frac{1}{2}\Gamma_n^{(r)}}{(E-E_r)+\frac{1}{2}i(\Gamma_n^{(r)}+\Gamma_a^{(r)})}\right|^2, \qquad \ldots\ldots(2)$$

where κ, E are the wave number and energy of the incident neutron and E_r is the energy which the neutron must have if it is to produce resonance in the compound nucleus. $\Gamma_n^{(r)}$ and $\Gamma_a^{(r)}$ are the widths of the resonance for re-emission of the neutron with its original energy and for absorption, respectively. E_r may be positive or negative but will numerically be much greater than the value of E for thermal neutrons, which is about 0·04 ev. With this assumption and other approximations eqn. (2) reduces to

$$\sigma=\frac{4\pi}{\kappa^2}\left|\kappa R-\frac{\frac{1}{2}\Gamma_n^{(r)}}{E_r}\right|^2=4\pi\left|R-\frac{\Gamma_n^{(r)}}{2\kappa E_r}\right|^2. \qquad \ldots\ldots(3)$$

It can be shown that the width $\Gamma_n^{(r)}$ is proportional to the wave number κ, so that eqn. (3) expresses the conclusion that the cross section σ is independent of the actual energy or wavelength of the neutrons, over the range of energies well removed from resonance levels. This is found to be in agreement with experiment. The quantity

$$a\equiv R-\Gamma_n^{(r)}/2\kappa E_r \qquad \ldots\ldots(4)$$

is termed the 'scattering length' of the nucleus. It is in fact equal to the amplitude of the scattered neutron beam at unit distance from the scattering nucleus, for an incident beam of unit amplitude.

By means of eqn. (4) we can interpret the general pattern of the variation of scattering length amongst the different nuclei. If $E_r/\Gamma_n^{(r)}$ is sufficiently large then a is approximately equal to R, giving the purely potential scattering which we have already mentioned. In this case, which is equivalent to scattering by an impenetrable sphere, the outgoing wave must be 180° out of phase with the incident wave, as indeed it is for the scattering of x-rays by electrons. If E_r becomes appreciably smaller then the value of a will fall below that of R and may indeed become negative. A negative value of a is in fact found for a few nuclei, namely H, ^{7}Li, Ti, V, Mn and ^{62}Ni. In the case of manganese, for example, this can be correlated with known scattering resonances at neutron energies of 300 ev and 2400 ev (Hibdon *et al.* 1950). On the other hand, chlorine has a scattering length which is much greater than R. This has been shown to be due to an energy level in the compound nucleus which corresponds to a negative value of E_r of -75 ev (Hibdon and Muehlhause 1950).

In general, then, nuclei have positive values of the scattering length a corresponding to a phase change of 180° on scattering. However, the few nuclei possessing negative a values, with 0° phase change, will be found to be of considerable interest in the practical applications of neutron diffraction. The conclusion that the scattering length is entirely real, with a phase change of either 0° or 180°, is not strictly correct and is due to neglect of the imaginary term in eqn. (2). This can be shown to be justifiable in practice in all cases except those of extremely high absorption. It amounts (Cassels 1950) to an assumption that

$$\sigma\lambda^2/\sigma_t^2 \gg \pi, \qquad \ldots\ldots(5)$$

where σ_t is the total cross section, equal to the sum of σ the scattering cross section and σ_a the absorption cross section. At a wavelength of 1 Å the factor on the left-hand side of (5) is of the order of 10^7 for most elements.

So far we have considered that the scattering nucleus is rigidly fixed and does not recoil. In the solid state this will usually be justifiable for the wavelengths used in diffraction work, except for the lightest nuclei, but it will not be so in the gaseous state, where the nuclei are free to move. It can be shown that the scattering cross section σ' for the free state is reduced below the value σ applicable to the bound state according to the relation

$$\sigma' = [A/(A+1)]^2\sigma, \qquad \ldots\ldots(6)$$

where A is the mass number of the nucleus. For hydrogen σ' is only a quarter of σ, but the difference between the bound and free cross sections rapidly becomes insignificant with increase of A.

A list of scattering lengths for various nuclei is given in table 1. Before drawing particular attention to this we emphasize two points. First, our discussion so far applies strictly only to nuclei which have zero spin. Secondly, we have been considering the behaviour of a single nuclear type, as distinct from the mixture of nuclei, or isotopes, which constitute most of the naturally occurring elements. We shall now consider the effect of these two factors.

Table 1. Neutron and X-Ray Scattering Data for the Elements

Element	Atomic number	Atomic weight	Specific nucleus	Nuclear spin	Neutrons			X-rays	
					$\bar{a}$ 10^{-13} cm	S 10^{-24} cm^2	σ 10^{-24} cm^2	f_x (10^{-13} cm) $\theta=0°$	$(\sin\theta)/\lambda = 0{\cdot}5$ Å^{-1}
H	1		^{1}H	1/2	−0·38[1]	1·8	81	0·28	0·02
			^{2}H	1	0·65	5·4	7·4	0·28	0·02
Li	3	6·94			−0·18	0·4		0·84	0·28
			^{6}Li	1	0·7	6		0·84	0·28
			^{7}Li	3/2	−0·25	0·8	2	0·84	0·28
Be	4		^{9}Be	3/2(?)	0·78	7·7	7·5	1·13	0·39
C	6		^{12}C	0	0·66	5·4	5·2	1·69	0·48
			^{13}C		0·60[2]	4·5	5·5	1·69	0·48
N	7		^{14}N	1	0·94[3]	11·1	10	1·97	0·53
O	8		^{16}O	0	0·58	4·2	4·2	2·25	0·62
F	9		^{19}F	1/2	0·55	3·8	3·5	2·53	0·75
Na	11		^{23}Na	3/2	0·35	1·5	3·5	3·09	1·14
Mg	12	24·3			0·54[4,5]	3·6	3·7	3·38	1·35
Al	13		^{27}Al	5/2	0·35	1·5	1·5	3·65	1·55
Si	14	28·06			0·40[6]	2·0		3·95	1·72
P	15	31·03			0·50[3]	3·1		4·23	1·83
S	16		^{32}S	0	0·31	1·2	1·2	4·5	1·9
Cl	17	35·5			0·99	12·2	15	4·8	2·0
K	19	39·1			0·35	1·5	2	5·3	2·2
Ca	20	40·1			0·49	3·0	3·5	5·6	2·4
			^{40}Ca	0	0·49	3·0	3·2	5·6	2·4
			^{44}Ca	0	0·18	0·4		5·6	2·4
Sc	21				1·02[8]	13·1		5·9	2·5
Ti	22	47·9			−0·38	1·8	6	6·2	2·7
V	23		^{51}V	7/2	−0·05[3]	0·03	5·1	6·5	2·8
Cr	24	52·0			0·35[7]	1·56	3·8	6·8	3·0
Mn	25		^{55}Mn	5/2	−0·37[7]	1·7	2·2	7·0	3·1
Fe	26	55·8			0·96	11·4	11·7	7·3	3·3
			^{54}Fe	0	0·42	2·2	2·5	7·3	3·3
			^{56}Fe	0	1·01	12·8	13	7·3	3·3
			^{57}Fe		0·23	0·64	2	7·3	3·3
Co	27		^{59}Co	7/2	0·28	1·0	5	7·6	3·4
Ni	28	58·7			1·03	13·4	18·0	7·9	3·6
			^{58}Ni	0	1·44[8]	25·9	27·0	7·9	3·6
			^{60}Ni	0	0·30[8]	1·1	1	7·9	3·6
			^{62}Ni	0	−0·87[8]	9·5	9	7·9	3·6
Cu	29	63·6			0·76	7·3	7·8	8·2	3·8
Zn	30	65·4			0·59	4·3	4·2	8·5	3·9
Ge	32	72·6			0·84	8·8	8·5	9·0	4·2
As	33		^{75}As	3/2	0·63	5·0	7	9·3	4·4
Se	34	79·0			0·89	10·0	10	9·6	4·5
Br	35	79·9			0·67	5·7	6·0	9·8	4·7
Rb	37	85·5			0·55	3·8	5·5	10·4	5·0
Sr	38	87·6			0·57	4·1	9·5	10·7	5·2
Zr	40	91·2			0·62	4·9	7	11·3	5·5
Cb	41		^{93}Cb	9/2	0·69[7]	6·0	6·2	11·5	5·7
Mo	42	95·9			0·66[7]	5·5	7·4	11·8	5·9
Pd	46	106·7			0·63	5·0	4·8	12·9	6·5

Table 1 (*cont.*)

Element	Atomic number	Atomic weight	Specific nucleus	Nuclear spin	Neutrons			X-rays	
					$\bar{a}$ 10^{-12} cm	S 10^{-24} cm^2	σ 10^{-24} cm^2	f_x (10^{-12} cm)	
								$\theta=0°$	$(\sin\theta)/\lambda$ $=0{\cdot}5$ Å^{-1}
Ag	47	107·9			0·61	4·6	7	13·3	6·7
			^{107}Ag	1/2	0·83	8·7	10	13·3	6·7
			^{109}Ag	1/2	0·43	2·3	6	13·3	6·7
Sn	50	118·7			0·61	4·6	4·9	14·1	7·2
Sb	51	121·8			0·54	3·7	4·2	14·4	7·4
Te	52	127·5			0·51[9]	3·3	3·9	14·7	7·6
I	53		^{127}I	5/2	0·52	3·4	3·8	15·0	7·7
Cs	55		^{133}Cs	7/2	0·49	3·0	7	15·5	8·1
Ba	56	137·4			0·53[8]	3·5		15·8	8·3
La	57	138·9			0·83[8]	8·7	9·3	16·1	8·4
Ce	58	140·25			0·46[8]	2·7	2·7	16·3	8·6
			^{140}Ce		0·47[8]	2·8	2·8	16·3	8·6
			^{142}Ce		0·45[8]	2·6	2·6	16·3	8·6
Pr	59	140·92			0·44[8]	2·4	4	16·6	8·8
Nd	60	144·27			0·72[8]	6·5	24	16·9	9·0
			^{142}Nd		0·77[8]	7·5		16·9	9·0
			^{144}Nd		0·28[8]	1·0		16·9	9·0
			^{146}Nd		0·87[8]	9·5		16·9	9·0
Ta	73		^{181}Ta	7/2	0·70	6·1	7·0	20·5	11·3
W	74	183·9			0·47[7]	2·7	6·8	20·8	11·4
Pt	78	195·2			0·95	11·2	11·2	22·0	12·1
Au	79		^{197}Au	3/2	0·76[7]	7·3	9	22·2	12·3
Hg	80				1·31[10]	21·5	26·5	22·5	12·5
Tl	81				0·75[11]	7·1		22·8	12·7
Pb	82	207·2			0·96[7]	11·5	11·6	23·1	12·9
Bi	83		^{209}Bi	9/2	0·89	10·1	10	23·3	13·1
Th	90		^{232}Th	0	1·01	12·8	12·8	25·2	14·4
U	92	238·1			0·85	9·0		25·8	14·7

For neutrons values are given of $\bar{a}$, the coherent scattering amplitude; S, the coherent scattering cross section; and σ, the total scattering cross section. The values are based on those of Shull and Wollan (1951), with later additions compiled by Dr. C. G. Shull from the sources indicated in the footnotes below :

[1] Burgy, M. T., Ringo, G. R., and Hughes, D. J., 1951, *Phys. Rev.*, **84**, 1160.
[2] Koehler, W. C., and Wollan, E. O., 1952, *Phys. Rev.*, **85**, 491.
[3] Peterson, S. W., and Levy, H. A., 1952, *Phys. Rev.*, **87**, 462.
[4] Bacon, G. E., 1952, *Acta Crystallogr.*, **5**, 684.
[5] Weiss, R. J. (unpublished).
[6] Milligan, W. O. (unpublished).
[7] Shull, C. G., and Wilkinson, M. K. (unpublished).
[8] Koehler, W. C., and Wollan, E. O. (unpublished).
[9] Heindl, C., and Ruderman, I. W., 1951, *Phys. Rev.*, **83**, 660.
[10] Hibdon, C. T., Muehlhause, C. O., Ringo, G. R., and Robillard, T. R., 1951, *Phys. Rev.*, **82**, 560.
[11] Winsberg, L., Meneghetti, D., and Sidhu, S. S., 1949, *Phys. Rev.*, **75**, 975.

The values of f_x, the x-ray scattering amplitudes, are given for the two cases of $(\sin\theta)/\lambda=0$ and $0{\cdot}5$ Å^{-1} (*International Tables for the Determination of Crystal Structures*, 1935, Vol. II).

(a) The effect of nuclear spin.

When the scattering nucleus has spin $\mathcal{I}$ it is possible for two different compound nuclei, having respectively spins of $\mathcal{I}+\frac{1}{2}$ and $\mathcal{I}-\frac{1}{2}$, to be formed by interaction with the neutron. Halpern and Johnson (1939) have shown that different scattering lengths a_+, a_- are applicable in the two cases. As a result of this the total scattering by such a nucleus is of two types. Part of the scattering is *coherent*, having a definite phase depending on the position of the nucleus, and is therefore able to produce interference effects with similar radiation from other nuclei, but the remainder is *incoherent*. The cross section σ is thus divided into a coherent part S and an incoherent part s, such that

$$\sigma = S + s. \qquad \ldots\ldots(7)$$

σ, S, s can be expressed in terms of a_+, a_- and p_+, p_-, where the latter two quantities are the relative weights which have to be ascribed to the two spin states of the compound nucleus. The relevant equations are

$$p_+ = (\mathcal{I}+1)/(2\mathcal{I}+1) \qquad \ldots\ldots(8) \qquad p_- = \mathcal{I}/(2\mathcal{I}+1) \qquad \ldots\ldots(9)$$

$$\sigma = 4\pi(p_+a_+^2 + p_-a_-^2) \qquad \ldots\ldots(10) \qquad S = 4\pi(p_+a_+ + p_-a_-)^2. \qquad \ldots\ldots(11)$$

If a_+, a_- are appreciably different, and particularly if they are of different sign, then S will be significantly smaller than σ and the difference is manifested in diffraction experiments as an increase in the background scattering. The coherent scattering amplitude of the nucleus is of course related to the coherent cross section S and is given by $\bar{a}$, where

$$\bar{a} = p_+a_+ + p_-a_-, \qquad \ldots\ldots(12)$$

instead of by the simple scattering length a for a spinless nucleus. The most striking example of the spin dependence of the scattering length is provided by hydrogen, for which a_+, a_- are of opposite sign. In contrast to the high σ value of 81×10^{-24} cm^2, S is only equal to $1{\cdot}8 \times 10^{-24}$ cm^2, and practically the whole of the scattering is incoherent.

In concluding this consideration of the effect of nuclear spin it is recalled that nuclei such as ^{12}C, whose mass number and atomic number are both even, have zero spin and accordingly no spin incoherence.

(b) The effect of isotopes.

A further complication arises when we take into account the fact that many naturally occurring elements consist not of a single nuclear species but of a number of different isotopes. The scattering lengths of different isotopes of an element may be different. As an example we may quote ^{58}Ni and ^{60}Ni, for which the values are 1·44 and $0{\cdot}30 \times 10^{-12}$ cm respectively.

In the three-dimensional crystal the various isotopic nuclei will be randomly distributed amongst the atomic sites allotted to an individual element. This gives rise to a further increase of background scattering, disordered scattering, at the expense of the coherent diffraction. Equations (10), (11) and (12) are generalized to give

$$\sigma = 4\pi\Sigma(p_r a_r^2) \qquad \ldots\ldots(13)$$

$$S = 4\pi(\Sigma p_r a_r)^2 \qquad \ldots\ldots(14)$$

$$\bar{a} = \Sigma(p_r a_r), \qquad \ldots\ldots(15)$$

where the summations are taken over the various possible isotopes and, where necessary, their positive and negative spin combinations. The factors p_r represent the probability of occurrence of any particular compound nucleus and are normalized so that Σp_r is equal to unity.

Table 1 which lists σ, S and $\bar{a}$ for various nuclei and elements may now be considered. These values are taken from Shull and Wollan's (1951) tabulation with later additions and modifications communicated by Dr. Shull. The methods by which they are derived by experimental measurement will be considered in a later section of this report. Perusal of the table will yield a number of examples, in addition to those already mentioned, of the effects of resonance, spin incoherence and isotope incoherence which we have discussed briefly above.

For convenience the values of the scattering amplitudes f_x for x-rays are also included in table 1. We shall draw attention subsequently to the comparative values of f_x, $\bar{a}$.

2.2. *Scattering by Microcrystals*

(*a*) *The total integrated scattering.*

In the foregoing discussion of scattering by nuclei and elements we have concentrated our attention on the quantity $\bar{a}$, the coherent scattering amplitude, since this will determine the intensities of the Bragg diffraction peaks from crystals which will be our main source of information concerning the structure of the solid state. Just as for x-rays, the intensity of any particular reflection depends on the unit cell 'structure factor'. This is given in terms of the atomic positions and scattering amplitudes by the well-known expression which appears below in eqn. (17) for the neutron case. It is however of some importance to consider also the summation of the number of neutrons, both coherent and incoherent, scattered in all directions, since experimental measurements of these quantities can be made and serve to determine some of the nuclear constants.

The total *coherent* scattering for a particular wavelength and temperature is defined as $E(S)$ and, for the case of a polycrystalline powder in which the microcrystals are oriented at random and are sufficiently small for absorption and extinction effects to be neglected, it has been shown by Weinstock (1944) and others that

$$E(S) = \tfrac{1}{2} N_c \lambda^2 \sum_{hkl} F_{hkl}^2 d_{hkl}. \qquad \ldots\ldots(16)$$

The summation is taken over all (hkl) planes for which Bragg reflections of the particular wavelength being used can take place, namely those planes for which the interplanar spacing d is greater than $\lambda/2$. N_c is the number of unit cells per cm^3 and F_{hkl} is the structure amplitude factor of the unit cell and is given by

$$F_{hkl}^2 = |\Sigma \bar{a} \exp\{2\pi i(hx/a_0 + ky/b_0 + lz/c_0)\}|^2, \qquad \ldots\ldots(17)$$

where the summation is taken over all the atomic positions (x, y, z) in the unit cell of dimensions a_0, b_0, c_0. Figure 1 shows the form of $E(S)$ as calculated for graphite from eqn. (16). Beyond the cut-off wavelength of 6·7 Å, which is twice the value of the (0002) interplanar spacing (this being the largest graphite spacing), $E(S)$ is equal to zero. At very short wavelengths it can be shown that the coherent scattering becomes isotropic and would be equal to S in the absence of effects due to inelastic scattering, which we shall consider shortly.

We define the total incoherent scattering by $E(s)$. This scattering is isotropic and it can be shown that the integrated value $E(s)$ is equal to s the incoherent scattering cross section which we have already defined by

$$s = \sigma - S$$
$$= 4\pi[\Sigma(p_r a_r^2) - (\Sigma p_r a_r)^2]. \quad \ldots\ldots(18)$$

(*b*) *Effect of inelastic collisions.*

In considering the interaction of neutrons with bound nuclei we assumed that the collisions were entirely elastic. In a three-dimensional crystal however this is only approximately true in general, for it is possible for energy to be lost by the neutron in the scattering collision. The energy is passed on to the crystal itself, thus setting the latter into oscillation in one of its permitted modes. Alternatively, it is also possible for the neutron to take up energy from the crystal in an inelastic collision. In the former case the scattered neutron will have a longer wavelength than that of the incident beam : in the latter case the neutron wavelength will be shortened.

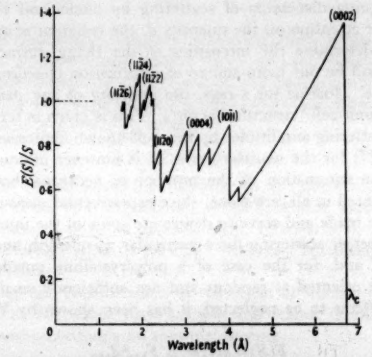


Fig. 1. The variation with wavelength of the total integrated scattering $E(S)$ for graphite, calculated for the assumption of no inelastic scattering. The discontinuities in the curve indicate the onset of reflection from the various crystallographic planes in the structure.

Weinstock (1944) was the first to calculate the effect of inelastic collisions, dealing with the case of a single isotope of zero nuclear spin. In Weinstock's treatment *elastic* scattering is regarded as a special case of inelastic scattering, and the theory deduces a variation of elastic scattering with temperature. The form and magnitude of this variation is exactly the same as that given by Debye (1913, 1914) and by Waller (1923, 1925) for x-ray scattering, according to the well-known exponential factor containing the Debye temperature. The intensities of the Bragg reflection peaks in neutron diffraction are thus corrected for inelastic scattering in precisely the same way as for x-ray diffraction.

In deducing an expression for the inelastic scattering Weinstock (1944) shows that for medium and long wavelengths the important transitions are those

in which a single quantum of energy is given to or taken from the crystal vibrations. Calculation in the particular case of iron at $\lambda = 1{\cdot}8$ Å shows that the inelastic scattering should be about 5% at room temperature, rising to 20% at 1000° K. Cassels (1951) has made an experimental check of the theory at 600° K using aluminium, which is free from the complicating effects of magnetic scattering which are found for iron, and concluded that the inelastic scattering is about 30% greater than is predicted from Weinstock's theory. It has since been shown by Squires (1952) that the difference can be accounted for by the effect of multi-oscillator processes in which there is a unit change of quantum number for more than one of the oscillators which may be considered to represent the various standing waves in the crystal. When correction is made for this Squires concludes that there is agreement with theory to an accuracy of better than 4% for both the wavelength and temperature variation of the inelastic scattering.

Cassels (1950) has extended the theory of Weinstock to the general case of elements having nuclei with finite spin and also consisting of a number of isotopes.

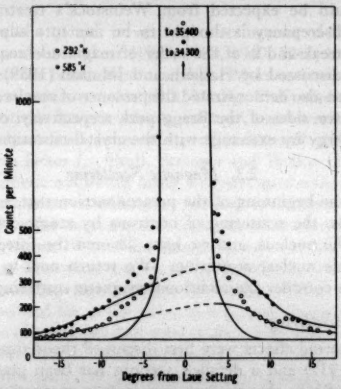


Fig. 2. The (110) diffuse streak for a single crystal of iron at two temperatures. The experimental intensities are resolved into the diffuse streak and the coherent Bragg reflection superimposed on a background of disordered scattering.

It is found that the total scattering is the sum of four terms, both the elastic E_{el} and inelastic scattering E_{in} being subdivided into ordered $E_{el}(S)$, $E_{in}(S)$ and disordered $E_{el}(s)$, $E_{in}(s)$ components. Expressions are given by Cassels (1950) for calculating the individual scattering terms from the known values of the crystal and nuclear constants. At long wavelengths where the neutron energy is small the probability of energy *emission* becomes zero and both of the inelastic terms become proportional to λ. It is in this long wavelength region where, as we have already noted, there is no ordered elastic scattering that experimental measurements to confirm the theory can be made most conveniently. Egelstaff (1951) has demonstrated experimentally the gain of energy when low energy neutrons are scattered. At short wavelengths the inelastic scattering rises rapidly again, with an appropriate decrease in the elastic scattering which

approaches zero on account of the exponential Debye factor, which has to be included in each of the terms of the summation in eqn. (16). However, the calculation at short wavelengths is complicated by the increasing importance of collisions in which the neutron gives several quanta of energy to the crystal vibrations. Finkelstein (1947) has calculated the behaviour in this region, employing a different approach from Weinstock's by using the Einstein model of the crystal vibrations, according to which all the thermal oscillators have a single frequency. Below a wavelength of about 0·2 Å the elastic scattering is negligible. Experimental measurements on beryllium by Fermi, Sturm and Sachs (1947) are in good agreement with Finkelstein's calculation of the total scattering.

Lowde (1952) has shown that diffuse streaks of inelastic scattering, such as are well known with x-rays (Lonsdale 1943), can also be observed with neutrons. Figure 2 shows the analysis of the diffracted beam from a single crystal of iron in the neighbourhood of the (110) reflection. The intensity of the diffuse streak is found to be closely proportional to the absolute temperature, as required by theory. In the case of iron the inelastic scattering is found to be considerably greater than would be expected from Weinstock's treatment of the *nuclear* scattering: the discrepancy is thought to be due to a superimposed *magnetic* inelastic diffuse streak and is of the order of magnitude required by the theory of this effect, as discussed by Halpern and Johnson (1939) and by Moorhouse (1951). Lowde has also demonstrated the presence of accelerated and decelerated neutrons on the two sides of the Bragg peak respectively, corresponding to the gain or loss of energy by exchange with the crystal vibrations.

2.3. *Magnetic Scattering*

We stated at the beginning of the present section that, except in the case of magnetic materials, the scattering of neutrons by atoms occurred by virtue of interaction with the nucleus, and we have devoted the foregoing paragraphs to a discussion of this nuclear scattering. We return now to the special case of magnetic atoms to consider the additional magnetic scattering which takes place.

(*a*) *Paramagnetic scattering.*

Magnetic scattering effects were first discussed theoretically by Bloch (1936) and Schwinger (1937) and a detailed account has been given by Halpern and Johnson (1939). Since the neutron possesses a magnetic moment, equal to 1·91 nuclear magnetons, it will interact with atoms such as the transition elements which have unbalanced electronic moments. In the case of paramagnetic materials the magnetic moments of atoms on individual sites are completely uncoupled and randomly directed. Under these circumstances Halpern and Johnson show that the magnetic scattering is completely incoherent. The cross section per atom per unit solid angle is

$$d\sigma_m = \tfrac{2}{3}\mathscr{S}(\mathscr{S}+1)(e^2\gamma/mc^2)^2 f^2, \qquad \ldots\ldots(19)$$

where $\mathscr{S}$ is the spin quantum number of the scattering atom, γ is the neutron magnetic moment expressed in nuclear magnetons and f the magnetic form factor which is dependent on $(\sin\theta)/\lambda$. The case where the ionic moment is due to spin only, with no orbital contribution, is being considered. The occurrence of the form factor f arises since the dimensions of the electronic orbits which contribute to the magnetic moment will be of the same order as the neutron wavelength. Consequently, although the contributions from the whole orbits

will be in phase at $\theta = 0°$, there will be increasing phase differences as the scattering angle is increased, giving a fall off from unity of the value of f. This magnetic form factor will be different from the usual x-ray form factor since it depends only on the few electron orbits which contribute to the magnetic moment. The total magnetic cross section per atom will be obtained by integrating eqn. (19) over all angles, bearing in mind the variation of f with angle.

The magnitude of $d\sigma_m$ can be illustrated for the case of the manganous ion Mn^{2+}. Here the ion is in the spectroscopic state $^6S_{5/2}$ with five electrons in the 3d shell having their spins all aligned parallel, giving a spin quantum number of 5/2. By substitution in eqn. (19) we find that $d\sigma_m = 1{\cdot}69f^2$ barn per unit solid angle. In the forward direction, therefore, where $f = 1$, $d\sigma_m = 1{\cdot}69$ barn, in comparison with a value of $\bar{a}^2$, which is the corresponding quantity for the *nuclear* scattering, of only 0·1 barn. In this case therefore the magnetic scattering at small angles is much greater than that from the nucleus. On the other hand, for Ni^{2+}, with a spin quantum number of 1, the value of $d\sigma_m$ is only 0·39 barn per unit solid angle in the forward direction in contrast with a much larger value of 1·06 barn for $\bar{a}^2$.

Early experimental attempts to verify the occurrence of this scattering were made by Whitaker and co-workers (1937, 1938, 1940, 1941) but were difficult in view of the weak neutron beams then available. More recently Ruderman (1949), making transmission measurements with MnF_2, MnO and α-Fe_2O_3 for slow monochromatic neutrons from a velocity selector, has correlated the additional scattering with Halpern and Johnson's predictions and deduced the angular variation of the form factor f. Shull, Strauser and Wollan (1951) have demonstrated the paramagnetic scattering using a crystal spectrometer. The effect of this scattering on the neutron diffraction patterns of compounds such as MnO, MnF_2 and $MnSO_4$ is to increase the background scattering. It was found experimentally that about 70% of the scattering near the forward direction still remained when due allowance was made by calculation for spin incoherence, thermal diffuse scattering and multiple scattering. The residual scattering, which is to be accounted for as the paramagnetic contribution, is readily placed on an absolute scale of barns per unit solid angle. The results of Shull, Strauser and Wollan are shown in fig. 3 for MnO, MnF_2 and $MnSO_4$. The curves for the latter two compounds show the expected angular variation and give a value of $d\sigma_m$ at $\theta = 0°$ which is about 10% lower than the calculated value of 1·69 barn. The curve for MnO suggests that at room temperature there is a certain amount of short range magnetic order rather than a completely random arrangement of the magnetic moments. The curve for MnF_2, which is considered the most reliable of those in fig. 3, can readily be converted into one showing the variation of amplitude form factor f with $(\sin\theta)/\lambda$. This curve is shown in fig. 4. From this latter curve Shull and his co-workers have calculated a radial distribution curve for the electrons in the 3d shell in Mn^{2+}. This distribution shows a maximum at a radius of about 0·6 Å and is illustrated in fig. 5, together with a curve calculated theoretically by Dancoff (see Shull, Strauser and Wollan 1951).

(*b*) *Ferromagnetic and antiferromagnetic scattering.*

Below a certain transition temperature the randomly oriented magnetic moments in a magnetic material may take up definite orientations. In the case of ferromagnetic materials all the magnetic moments within a single domain

will become parallel, although in the absence of a magnetic field there will be no correlation in the orientations from domain to domain. Equally, in antiferromagnetic materials the magnetic atoms may be considered to lie on two interpenetrating lattices which, below the transition temperature, have their occupying atoms arranged with their moments antiparallel. In either of these cases, and indeed in the intermediate case of ferrimagnetism which is met in

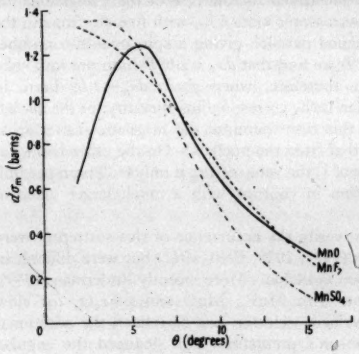


Fig. 3. The differential magnetic scattering cross sections for several paramagnetic salts containing Mn^{2+} ions. (*After* Shull, Strauser and Wollan 1951)

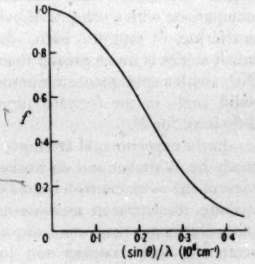


Fig. 4. Variation of magnetic amplitude form factor of Mn^{2+} with $(\sin\theta)/\lambda$. (*After* Shull, Strauser and Wollan 1951)

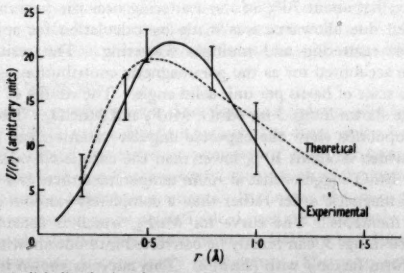


Fig. 5. Electron radial distribution function for the 3d shell of Mn^{2+}, with comparison of theoretical and experimentally deduced curves. The vertical lines indicate possible experimental error. (*After* Shull, Strauser and Wollan 1951)

the ferrites, coherence appears between the magnetic scattering amplitudes of the various atoms, now that the magnetic moments of these have become arranged in ordered, defined orientations.

Halpern and Johnson (1939) have shown that to each magnetic atom under these circumstances may be ascribed a magnetic scattering amplitude D, which can be treated in corresponding fashion to the nuclear scattering amplitude $\bar{a}$.

Coherent Bragg diffraction peaks depending on the three-dimensional crystal structure will be built up as for nuclear scattering, the domains behaving as crystallites, and if the magnetic unit cell is of the same dimensions as the chemical unit cell then these magnetic Bragg peaks will coincide in position with the ordinary nuclear peaks. In some cases, as will be seen in a later section, the magnetic unit cell may have dimensions which are twice those of the chemical unit cell, and under these conditions magnetic peaks will appear in new positions corresponding to the super-lattice. However, there is an important conclusion from Halpern and Johnson's work which must be emphasized here. When the neutron beam is *unpolarized* there is no coherence between the nuclear and magnetic scattering, and it is the intensities, and not the amplitudes, which are additive when nuclear and magnetic peaks occur at the same position.

(*c*) *Unpolarized neutron beams.*

The significance of the proviso that the neutron beam should be unpolarized may be seen from Halpern and Johnson's general expression. According to this the differential scattering cross section $d\sigma_m$ of a magnetic ion is

$$d\sigma_m = \bar{a}^2 + 2\bar{a}D\mathbf{q}.\boldsymbol{\lambda} + D^2q^2, \qquad \ldots\ldots(20)$$

where $\bar{a}$ is the nuclear scattering amplitude, D is the magnetic scattering amplitude, $\boldsymbol{\lambda}$ is a unit vector in the direction of polarization of the incident neutron, $\mathbf{q}$ is the magnetic interaction vector, equal to $\boldsymbol{\epsilon}(\boldsymbol{\epsilon}.\mathbf{K})-\mathbf{K}$, where $\boldsymbol{\epsilon}$ is the unit scattering vector, and $\mathbf{K}$ is a unit vector in the direction of the magnetic moment of the ion. It follows that

$$q^2 = 1-(\boldsymbol{\epsilon}.\mathbf{K})^2 \qquad \ldots\ldots(21)$$

$$= \sin^2\alpha, \qquad \ldots\ldots(22)$$

where α is the angle between the scattering and magnetization vectors.

If the incident radiation is unpolarized then the average value of $\mathbf{q}.\boldsymbol{\lambda}$ is zero and eqn. (20) reduces to

$$d\sigma_m = \bar{a}^2 + D^2q^2, \qquad \ldots\ldots(23)$$

expressing the fact that the *intensities* of the nuclear scattering and the magnetic scattering are additive.

The numerical value of q^2, which is given by eqn. (21), depends on the relative orientation of the magnetic moment and the scattering vector. If the magnetic moment vector is perpendicular to the scattering planes being considered (i.e. it is parallel to the scattering vector) then $\boldsymbol{\epsilon}.\mathbf{K}=1$ and q is zero, whereas if the magnetic vector lies in the planes then $q=1$. When the magnetic moment is directed randomly with respect to the crystallographic axes, although in alignment from atom to atom in a domain, then q^2 has an effective mean value of $\frac{2}{3}$. The value is also $\frac{2}{3}$ when, as in the case of iron, the direction of common alignment within a domain may be *any one* of the cube axes.

In contrast with the nuclear scattering amplitude, which at present has to be determined empirically by experiment, the magnetic scattering amplitude can be calculated from spectroscopic data. Halpern and Johnson (1939) show that

$$D = (e^2\gamma/mc^2)\mathscr{S}f, \qquad \ldots\ldots(24)$$

where the symbols have the same significance as given for eqn. (19). It follows that

$$D = 0{\cdot}539\times10^{-12}\mathscr{S}f \text{ cm.} \qquad \ldots\ldots(25)$$

For example, the value of D for Fe^{3+}, which has a spin quantum number of 5/2, is equal to $1{\cdot}360f \times 10^{-12}$ cm: for metallic iron, with an effective spin quantum number of 1·11 determined by measurement of saturation magnetization, D is equal to $0{\cdot}598f \times 10^{-12}$ cm. For Fe^{3+} the form factor arrived at above for Mn^{2+} is a sufficiently good approximation: for metallic iron a curve calculated by Steinberger and Wick (1949) is shown by Shull, Wollan and Koehler (1951) to be in good agreement with experiment.

We may summarize by saying that for unpolarized incident neutrons the nuclear and magnetic intensities are calculated separately, using the nuclear scattering amplitude $\bar{a}$ and the magnetic scattering amplitude D, containing the form factor applicable to the angle θ, with the appropriate value of q^2. This procedure will be illustrated later when studies of antiferromagnetic and magnetized ferromagnetic materials are described.

(*d*) *Scattering by magnetized samples.*

The effect of the factor q^2 in determining the magnetic intensity can be seen directly from certain experiments with magnetic fields. The most favourable case, insensitive to depolarization effects which will be discussed later, is provided by the (111) reflection of Fe_3O_4. Shull, Wollan and Koehler (1951) have shown that whereas in the unmagnetized condition the magnetic scattering intensity is proportional to $F_{nucl}^2 + \frac{2}{3}F_{mag}^2$, on magnetization it will approach F_{nucl}^2 or $F_{nucl}^2 + F_{mag}^2$ respectively for the two cases in which the magnetic field is parallel or perpendicular to the scattering vector. The experimental results for these and intermediate field directions are in very good agreement with Halpern and Johnson's theory. F_{nucl}, F_{mag} are the structure factors for the unit cell using the nuclear and magnetic amplitudes $\bar{a}$, D respectively.

2.4. *Polarization Effects in Magnetized Samples*

(*a*) *Polarization of diffracted beams.*

So far we have considered the mean magnetic scattering intensity averaged over all the spin orientations in an unpolarized neutron beam. If we consider such a beam incident on a magnetized sample it may be resolved into two equal and independent components having their spins parallel and antiparallel to the magnetic field. For the particular case where the field direction is perpendicular to the scattering vector $\boldsymbol{\epsilon}$ it follows that $\boldsymbol{\epsilon}.\mathbf{K}$ is zero, $\mathbf{q} = -\mathbf{K}$ and $q^2 = 1$. Thus eqn. (20) reduces to $d\sigma_m = \bar{a}^2 \pm 2\bar{a}D + D^2$, the $\pm$ sign depending on whether the spin component parallel or antiparallel to the field is being considered. Hence

$$d\sigma_m = (\bar{a} \pm D)^2. \qquad \ldots\ldots(26)$$

In this case therefore coherence is achieved between the nuclear and magnetic scattering. Moreover, if we can choose a reflection and material for which $\bar{a}$, D are approximately equal then the diffracted beam will be effectively completely polarized. Shull, Wollan and Koehler (1951) have demonstrated this for the case of the (220) reflection of Fe_3O_4. The structure factor for this reflection is $(8a_{Fe})^2$, where a_{Fe} is the scattering amplitude for an Fe^{3+} ion, and is not dependent on the scattering by oxygen. The nuclear scattering amplitude $\bar{a}$ for iron is $0{\cdot}96 \times 10^{-12}$ cm, and this is also the value of D, allowing for the form factor, for this reflection. Consequently $\bar{a} - D$ vanishes and the diffracted beam is completely polarized within the experimental uncertainty of, say, 5%. This method allows

the attainment of a collimated beam of monochromatic, polarized neutrons with an intensity of about 10^5 neutrons/second. If required, the beam can be pulsed (Shull, Wollan and Strauser 1951).

(*b*) *Polarization by transmission.*

The above method whereby complete polarization of the Bragg reflected beam from a suitable magnetized single crystal is produced is to be contrasted with the older method of polarization by transmission through a polycrystalline sample. When a neutron beam is incident on a magnetized specimen of polycrystalline iron diffraction will take place into the various Debye–Scherrer rings, giving intensities which are determined, for any individual direction of scattering, by eqn. (20). It follows from the presence of the term $\mathbf{q}.\boldsymbol{\lambda}$ in this equation that $d\sigma_m$ will be different for the components of the incident beam having their spins parallel and antiparallel to the magnetic field. Consequently the effective cross section σ_{eff} for a transmission experiment, which will show a similar peaked variation to that discussed for nuclear scattering in fig. 1, will be different for the two spin orientations. It was originally shown by Bloch (1936, 1937) that for a magnetically saturated material

$$\sigma_{eff}=\sigma'_{eff}\pm\sigma_p, \qquad \ldots\ldots(27)$$

where σ'_{eff} is the cross section for complete demagnetization and σ_p is the polarization cross section. One polarization state is preferentially scattered out of the beam and there is a net *increase* ΔI in the transmitted intensity, on magnetization, given by

$$\Delta I/I=\cosh(N\sigma_p t)-1=\tfrac{1}{2}N^2\sigma_p{}^2t^2 \quad \text{for} \quad N\sigma_p t\gg 1, \qquad \ldots\ldots(28)$$

where I is the transmitted intensity in the demagnetized condition, N is the number of nuclei per cm^3 and t is the sample thickness. Thus σ_p can be measured experimentally, and has been shown by Hughes, Wallace and Holtzman (1948) and by Fleeman, Nicodemus and Staub (1949) to have the expected discontinuous variation with λ. Good quantitative agreement is obtained with the most recent calculation, by Steinberger and Wick (1949) on the lines of the earlier work by Halpern, Hamermesh and Johnson (1941) and Hamermesh (1942). For a beam of neutrons from the thermal column of a pile σ_p is about 3 barns.

A high degree of polarization, which can be measured by passing through a second analysing block of iron, is only achieved asymptotically with increasing thickness of the iron sample. In practice with a field of 15000 gauss, a polarization of about 70% can be obtained. If the iron is not magnetized to saturation the inhomogeneities in the internal field cause depolarization (Halpern and Holstein 1941, Fleeman, Nicodemus and Staub 1949).

(*c*) *Polarization by reflection from mirrors.*

More recently Hughes and Burgy (1951) have studied the production of polarization by reflection from magnetized mirrors. They have demonstrated that for a magnetic material there are two refractive indices, corresponding to the two spin states, given by

$$n^2=1-\lambda^2N\bar{a}/\pi\pm\gamma B_s/E, \qquad \ldots\ldots(29)$$

where γ is the neutron magnetic moment, E the neutron energy and B_s is the

magnetic induction in a domain. The magnetic interaction term can be written alternatively as $\pm\lambda^2 ND/\pi$, where D is the magnetic scattering amplitude in the forward direction derived from eqn. (24). In the case of iron, for example, the D value of $0{\cdot}598\times10^{-12}$ cm (Shull, Wollan and Koehler 1951) is equivalent to a value of B_s of 22 500 gauss. The actual form of this magnetic term has been the subject of much theoretical discussion (Halpern, Hamermesh and Johnson 1941, Halpern 1949, Ekstein 1949, 1950, Lax 1950) which is considered by Hughes and Burgy (1951) in relation to their experimental results.

Corresponding to the two values of n there are two different critical angles for the two spin states, given by

$$\theta_c=[\lambda^2 N\bar{a}/\pi\pm\gamma B_s/E]^{1/2}. \qquad \ldots\ldots(30)$$

For a cobalt mirror the magnetic term is greater than the nuclear term, so that the refractive index is less than unity only for one of the two spin states. Thus for saturated cobalt complete polarization can be produced in the totally reflected beam without monochromatizing the incident radiation. This has been achieved by Hughes and Burgy to an accuracy of within 1%, giving a fully polarized beam of 10^3–10^4 neutrons per second. For iron, on the other hand, the nuclear amplitude is the larger, and only incomplete polarization can be obtained for an incident 'white' beam, corresponding to total reflection of a larger range of λ for one spin state than for the other. However, complete polarization would be expected for a monochromatic beam incident on a magnetized iron mirror at an angle lying between the two critical values given by eqn. (30).

§3. DIFFRACTION BY CRYSTALS

The discussion of the previous section was concerned with the diffraction of neutrons by individual nuclei, unit cells of crystal and small fragments of polycrystal in which it could be assumed that any absorption or attenuation of the incident beam in traversing the crystallite could be neglected. We proceed now to consider the diffraction of neutrons by single crystals of significant size. Our investigation will follow the general lines which are pursued in considering the same problem in x-ray diffraction. We shall find that the physical principles involved are the same in the two cases but that different approximations are permitted in typical practical examples, resulting in conclusions which are significantly different.

We first call attention to the absolute magnitudes of f_x, $\bar{a}$ which are the coherent scattering amplitudes for x-rays and neutrons respectively, as listed earlier in table 1 for many of the elements. For hydrogen and deuterium the neutron scattering is greater, but on account of the steady increase of x-ray scattering with atomic number the f_x-values quickly exceed $\bar{a}$. However, it is not until an atomic number of about 50 is reached that f_x consistently exceeds $\bar{a}$ by a factor of 10. It follows therefore that for many substances the ratios of the amplitudes of reflected and incident beams, when neutrons or x-rays are diffracted by nuclei or unit cells in a crystal, will be the same within a factor of 10. Consequently the phenomenon of primary extinction (Darwin 1914) and the concepts of 'perfect' and 'mosaic' crystals (Lonsdale 1947), which arise in the discussion of x-ray intensity determinations, will be significant in the neutron case also.

Calculation, in a similar way to that described by James (1948), shows that the amplitude reflected by a single net plane of a crystal will be of the order of

a factor 10^{-4} of the incident amplitude for X-rays and a little less for neutrons. In both cases 'perfect' crystals will give integrated reflections which initially increase with increasing crystal thickness but rapidly reach a constant value when t exceeds more than 10^4 planes. This behaviour is illustrated in fig. 6, for which the ordinate is the integrated reflection $\mathscr{R}^0$ divided by ζ, the half-width of the angular range over which the integrated reflection of a perfect crystal is unity. It can be shown that

$$\zeta = \lambda^2 N_e F/\pi \sin 2\theta, \qquad \ldots\ldots(31)$$

which amounts to less than a second of arc for neutrons and a few seconds for X-rays.

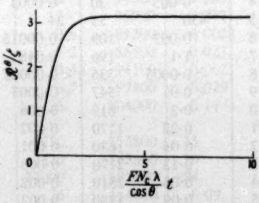


Fig. 6. Variation of the integrated reflection of a perfect crystal slab with thickness t for the 'reflection' case.

We come to a significant difference for neutrons when we examine the reflection by a 'mosaic' crystal. Here the atomic arrangement is no longer perfect throughout the body of the crystal but is interrupted by occasional faults and dislocations which occur at separations of the order of 5000 Å. Such a crystal is usually regarded as being built up of a mosaic of small blocks, within which the structure is perfect but which are slightly misoriented, to an extent generally equal to a few seconds of arc, among themselves. When a beam of radiation is reflected by a crystal of this sort, oriented in its Bragg reflection position, the perfect shielding of the lower planes by the upper ones does not occur. If the mosaic blocks are appreciably less than 10^4 planes in thickness then the saturation limitation shown in fig. 6 will not occur and the integrated reflection will continue to increase with crystal thickness. Eventually, however, the beam will reach a depth in the crystal at which shielding of *block* by *block* becomes significant and the reflection becomes limited by what is called 'secondary extinction'.

With X-rays secondary extinction does not have any major effect in determining the integrated reflection, since the beam is attenuated much more rapidly by true absorption, giving photoelectric emission and fluorescent scattering, than by the scattering into the diffracted beam. Thus the beam as a whole never penetrates far enough for secondary extinction to be predominant. On the other hand the coefficients of true absorption, by capture processes, for neutrons are very much smaller and it is indeed the scattering process which primarily reduces the intensity of a neutron beam when it passes through a crystal. The entirely different orders of magnitude of the two sets of absorption coefficients will be clear from table 2, which lists the linear absorption coefficients of some of the solid elements for neutrons ($\lambda = 1{\cdot}08$ Å) and X-rays ($\lambda = 1{\cdot}54$ Å). In the exceptional case of boron

Table 2. Absorption Data for Elements for Neutrons ($\lambda = 1{\cdot}08$ Å) and X-Rays ($\lambda = 1{\cdot}54$ Å)

Element		Atomic number	σ_a 10^{-24} cm^2		μ/ρ cm^{-1}		μ cm^{-1}	
			Neutrons	X-rays	Neutrons	X-rays	Neutrons	X-rays
Hydrogen	H	1	0·19		0·11			
Helium	He	2						
Lithium	Li	3	40		3·5		1·87	
Beryllium	Be	4	0·005	20	0·0003	1·3	0·00054	2·3
Boron	B	5	430	55	24	3·1	60	7·7
Carbon	C	6	0·003	109	0·00015	5·5	0·0005	19·2
Nitrogen	N	7	1·1	196	0·048	8·5		
Oxygen	O	8	< 0·0005	336	< 0·00002	12·7		
Fluorine	F	9	< 0·01	547	< 0·0003	17·5		
Neon	Ne	10	0·2	819	0·006	24·6		
Sodium	Na	11	0·28	1170	0·007	30·9	0·007	30
Magnesium	Mg	12	0·04	1630	0·001	40·6	0·0017	70
Aluminium	Al	13	0·13	2170	0·003	48·7	0·008	131
Silicon	Si	14	0·06	2810	0·002	60·3	0·004	145
Phosphorus	P	15	0·09	3730	0·002	73·0	0·004	146
Sulphur	S	16	0·28	4820	0·0055	91·3	0·011	182
Chlorine	Cl	17	19·5	5040	0·33	103		
Argon	A	18	0·4	7420	0·0060	113		
Potassium	K	19	1·2	9230	0·018	143	0·016	125
Calcium	Ca	20	0·25	11380	0·0037	172	0·0057	266
Scandium	Sc	21	7·1	13800	0·09	185	0·28	555
Titanium	Ti	22	3·5	16160	0·044	204	0·20	920
Vanadium	V	23	2·8	19100	0·033	227	0·20	1360
Chromium	Cr	24	1·8	22200	0·021	259	0·15	1840
Manganese	Mn	25	7·6	23700	0·083	284	0·60	2040
Iron	Fe	26	1·4	29800	0·015	324	0·12	2570
Cobalt	Co	27	21	34300	0·21	354	1·87	3160
Nickel	Ni	28	2·7	4750	0·028	49·2	0·25	438
Copper	Cu	29	2·2	5620	0·021	52·7	0·19	470
Zinc	Zn	30	0·6	6350	0·0055	59·0	0·039	420
Gallium	Ga	31	1·8	7310	0·015	63·3	0·089	374
Germanium	Ge	32	1·3	8300	0·011	69·4	0·058	374
Arsenic	As	33	2·5	9450	0·020	76·5	0·12	436
Selenium	Se	34	7·4	10700	0·056	82·8	0·27	398
Bromine	Br	35	3·8	12020	0·029	92·6		
Krypton	Kr	36	0·027	13600	0·0002	100		
Rubidium	Rb	37	0·42	15400	0·0029	109	0·0044	167
Strontium	Sr	38	0·7	17100	0·0048	119	0·012	303
Yttrium	Y	39	0·83	19000	0·0056	129	0·031	710
Zirconium	Zr	40	0·10	21200	0·0006	143	0·0041	915
Niobium	Nb	41	0·63	23400	0·0041	153	0·034	1290
Molybdenum	Mo	42	1·4	26000	0·009	164	0·08	1670
Technicium	Tc	43						
Ruthenium	Ru	44	1·5	30900	0·009	185	0·10	2250
Rhodium	Rh	45	90	33600	0·53	198	6·6	2480
Palladium	Pd	46	4·0	36400	0·023	207	0·28	2530
Silver	Ag	47	36·0	39600	0·20	223	2·0	2340
Cadmium	Cd	48	2100	43400	11·2	234	97	2020
Indium	In	49	115	47600	0·6	252	4·4	1840
Tin	Sn	50	0·35	51600	0·002	265	0·011	1540

Table 2 (*cont.*)

Element		Atomic number	σ_a 10^{-24} cm²		μ/ρ cm^{-1}		μ cm^{-1}	
			Neutrons	X-rays	Neutrons	X-rays	Neutrons	X-rays
Antimony	Sb	51	3·2	56900	0·016	284	0·10	1910
Tellurium	Te	52	2·7	60600	0·013	289	0·081	1800
Iodine	I	53	3·7	65600	0·018	314	0·09	1540
Xenon	Xe	54	18	71000	0·083	330		
Caesium	Cs	55	17	75900	0·077	347	0·15	660
Barium	Ba	56	0·6	81300	0·0027	359	0·01	1260
Lanthanum	La	57	5·3	86500	0·023	378	0·14	2340
Cerium	Ce	58	0·48	94100	0·0021	407	0·015	2810
Praseodymium	Pr	59	6·7	97800	0·029	422	0·19	2740
Neodymium	Nd	60	26	104000	0·11	437	0·75	3020
Prometheum	Pm	61						
Samarium	Sm	62	6350	115800	25	467	190	3670
Europium	Eu	63	2520	115600	10	461		
Gadolinium	Gd	64	22000	122000	84	470		
Terbium	Tb	65	26	114000	0·09	435		
Dysprosium	Dy	66	535	124000	2·0	462		
Holmium	Ho	67	40	34500	0·15	128		
Erbium	Er	68	100	36900	0·36	133		
Thulium	Tm	69	71	38600	0·25	139		
Ytterbium	Yb	70	22	41000	0·076	144		
Lutecium	Lu	71	65	43600	0·22	151		
Hafnium	Hf	72	61	46100	0·20	157	2·7	2080
Tantalum	Ta	73	13	49000	0·044	164	0·73	2720
Tungsten	W	74	11	51800	0·036	171	0·70	3300
Rhenium	Re	75	50		0·16		3·3	
Osmium	Os	76	9	58500	0·028	186	0·63	4190
Iridium	Ir	77	260	61800	0·80	194	18	4360
Platinum	Pt	78	5	66000	0·02	205	0·36	4400
Gold	Au	79	57	69500	0·17	214	3·3	4140
Mercury	Hg	80	210	73600	0·63	223	8·4	3020
Thallium	Tl	81	2·0	77600	0·006	231	0·07	2720
Lead	Pb	82	0·1	82200	0·0003	241	0·003	2720
Bismuth	Bi	83	0·01	86800	0·00003	253	0·0003	2870
Polonium	Po	84						
Astaline	At	85						
Radon	Rn	86		102000		278		
Francium	Fr	87						
Radium	Ra	88		113400		304		
Actinium	Ac	89						
Thorium	Th	90		125000		327		3690
Protoactinium	Pa	91						
Uranium	U	92	2·1	138000	0·005	352		6600

σ_a = cross section for true absorption (capture).
μ/ρ = mass absorption coefficient.
μ = linear absorption coefficient of the solid element.
Neutron data derived from *Nuclear Data* (1950) and Pomerance (1951, 1952).

μ for neutrons is greater than for x-rays, but in all other cases, even those of cadmium and some of the rare earths, which are regarded as very effective absorbers of thermal neutrons, the x-ray absorptions are very much greater.

The effect of the very low neutron absorption coefficients on single crystal intensities was first pointed out by Bacon and Thewlis (1949), who showed that for thick crystals the integrated reflection would be largely determined by the angular spread of the mosaic blocks and little by the unit cell structure factor. A detailed investigation was described by Bacon and Lowde (1948), who extended the treatment given by Zachariasen (1945) for the x-ray case to allow for the effect of very small absorption. These authors treated the reflection of neutron beams by large mosaic slabs in both the 'transmission' and 'reflection' positions for both symmetrical and asymmetrical conditions, and presented curves whereby the integrated reflection could be obtained in practical cases. For details the reader is referred to the original paper, but attention will be drawn to the main conclusions, which can be derived from the curves shown in fig. 7 for the symmetrical reflection case, in which the reflecting planes are parallel to the crystal surface.

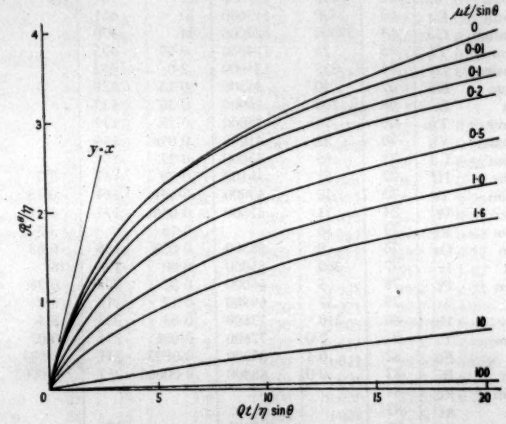


Fig. 7. The integrated reflection for an ideally imperfect crystal ('reflection' case).

The choice of the coordinates and parameters for the curves follows from the analysis, which derives an expression for the integrated reflection $\mathscr{R}^0$ as a function ψ of the form

$$\frac{1}{\eta}\mathscr{R}^0 = \psi\left(\frac{\mu t}{\sin\theta}, \frac{1}{\eta}\frac{Q}{\mu}\right), \qquad \ldots\ldots(32)$$

where η is the standard deviation of the angular spread of the mosaic blocks, which are assumed to have a gaussian distribution curve, μ is the linear coefficient of true absorption, t is the thickness of the crystal, θ is the Bragg angle, Q is the crystallographic quantity given by

$$Q = N_c^2\lambda^3 F^2/\sin 2\theta, \qquad \ldots\ldots(33)$$

where N_c is the number of unit cells per cm^3. F is the structure factor of the unit cell for neutrons, as defined in eqn. (17), with addition of the appropriate Debye exponential function to correct for thermal vibrations.

We note the following features of the curves:

(*a*) As $Qt/\eta \sin\theta$ increases there is an initial region near the origin where $\mathscr{R}^0/\eta$ is *equal* to $Qt/\eta \sin\theta$, i.e. the integrated reflection is given by QV where V is the total volume of crystal irradiated by the beam. This initial part of the curves coincides with the line $y=x$, but is too small to be distinguished in the lowest curves.

(*b*) For the curves with large values of $\mu t/\sin\theta$, which correspond to the case of x-ray diffraction, $\mathscr{R}^0/\eta$ subsequently becomes *proportional* to $Qt/\eta \sin\theta$ and, in fact, $\mathscr{R}^0$ is equal to $Q/2\mu$, as is well known.

(*c*) For the curves with small values of $\mu t/\sin\theta$, which are the ones in general applicable to neutron diffraction, the fall off from the line $y=x$ is followed by a region where $\mathscr{R}^0$ is no longer proportional to Q but is determined largely by the mosaic spread η of the crystal. Over this region secondary extinction is predominant and the value of $\mathscr{R}^0$ is of the order of 3η.

In the intermediate region the curves are approximately parabolic in form, and Bacon and Lowde (1948) state that, when $\mu=0$, $\mathscr{R}^0$ is given to an accuracy of within 5% by the parabola

$$\mathscr{R}^0 = 0{\cdot}96(\eta Qt/\sin\theta)^{1/2} \qquad \ldots\ldots(34)$$

for values of $Qt/\eta \sin\theta$ ranging from 1·9 to 13.

3.1. *Experimental Measurements with Single Crystals*

An experimental verification of the above theory has been given by Bacon (1951), who describes measurements of the integrated reflections for the (200), (400), (600) spectra of potassium bromide, using crystal slabs varying from 0·17 to 1·38 cm in thickness. The measurements cover values of $\mu t/\sin\theta$ ranging from 0·06 to 1·45 and values of $Qt/\eta \sin\theta$ from 0·04 to 50, so that conditions range from the region where the QV law is applicable to well beyond the confines of fig. 7, where secondary extinction is predominant. It is concluded that the experimental intensities, on an absolute basis, are within 10% of those calculated theoretically throughout the range of the experiment.

Prior to this work some earlier intensity measurements with large single crystals were reported by Fermi and Marshall (1947) in the course of their fundamental determinations of the sign of the scattering lengths of some of the elements. These workers noted that their reflectivities were more nearly proportional to the structure factor F than to its square, or to Q, as they expected for their mosaic crystals by analogy with x-rays. This behaviour can be interpreted readily by concluding that they were working in the intermediate region of partial secondary extinction which is covered by the approximate parabolic relation of eqn. (34).

From a consideration of the initial portion of the curves such as are given in fig. 7 Bacon and Lowde (1948) deduced a criterion for the maximum thickness of crystal for which secondary extinction can be neglected for neutrons. They conclude that $\mathscr{R}^0$ will be within 5% of the value QV if

$$Qt/\eta \sin\theta < \tfrac{1}{4} \qquad \ldots\ldots(35)$$

for the 'reflection' case and if

$$Qt/\eta \cos\theta < \tfrac{1}{4} \qquad \ldots\ldots(36)$$

for the 'transmission' case. These criteria offer a guide to deciding on the maximum size of crystal which will permit proportionality between $\mathscr{R}^0$ and Q to be assumed, thus enabling intensity measurements to be interpreted easily without making what is often an uncertain correction for secondary extinction, particularly in view of the difficulty of arriving at a precise knowledge of the mosaic spread η of an individual crystal.

Although plane parallel slabs of crystal are convenient for testing the accuracy of the extinction theory they are not convenient for structure analysis purposes, being not only difficult to obtain but also awkward to use for planes not parallel to the surface. Ideally a small spherical crystal is the most convenient, but at the present time the neutron beam intensities are too weak to permit the use of a sphere of radius sufficiently small to be free from extinction errors in all directions. A more convenient form of crystal is a cylindrical pillar, mounted with its axis perpendicular to the neutron beam, with which it is possible to measure all reflections in the zone having its axis parallel to the cylinder axis. In this case it would be expected from eqns. (35), (36) that the criterion for agreement with QV to within 5% would be

$$R_s < \eta/8Q, \qquad \ldots\ldots(37)$$

where R_s is the radius of the cylinder, or the mean transverse dimension of a square or hexagonal crystal, which is often more convenient in practice.

Some measurements by Bacon (1951) with square pillars of potassium bromide demonstrate the truth of eqn. (37). In this particular example, for which the maximum Q value used was $0{\cdot}0010\,\mathrm{cm}^{-1}$, for the 200 reflection, it was found that for $\eta = 5'$ and $R_s = 1\,\mathrm{mm}$ satisfactory proportionality between $\mathscr{R}^0$ and Q was obtained. On the other hand, for $\eta = 1\frac{1}{2}'$ and $R_s = 2\,\mathrm{mm}$ there was appreciable secondary extinction, although for this specimen the extinction could be made practically negligible by dipping the crystal in liquid air to increase the mosaic spread.

In practice a wide range of values of Q, η is found and individual calculation is necessary to decide whether for any particular substance a crystal large enough to give adequate neutron counting rates will be small enough to be reasonably free from extinction. In many cases, as pointed out by Bacon (1951) and by Peterson and Levy (1951), this will certainly be so, but in other cases extinction will be too severe. This is illustrated by the latter authors who, for various substances, compare the structure factors deduced from experimental measurements with single crystal and powdered specimens respectively. These results are shown in table 3.

3.2. *The Use of Powder Methods*

An obvious method of avoiding the difficulties involved in correlating single crystal intensities with Q values is to use the powder diffraction technique, analogous to that associated with the names of Debye and Scherrer for x-rays, in which the material under examination is ground into fragments so small that extinction effects are negligible. In many cases, of course, this technique may be the only one available, for many substances will be unobtainable in single crystals of the order of millimetres in dimension which would certainly be required for adequate neutron counting rates. In fact, only two of the structural investigations so far reported

Table 3. Comparison of Structure Factors determined Experimentally from Single Crystal and Powder Diffraction Measurements (Peterson and Levy 1951)

Substance	Reflection	Structure factor $\times 10^{12}$ cm			Thickness (mm)	
		Observed (powder)	Observed (single crystal)	Calculated	Thin crystal limit for $\eta = 5'$	Measured value
KBr	111	1·29	1·27	1·28		
	200	4·04	4·09	4·08	1·6	1·9
KCl	111	2·40	2·37	2·56		
	200	5·04	4·71	5·36	0·7	1·8
KHF_2	121	3·09	3·08	3·06	2·3	2·8
NH_4Cl	110	1·34	1·27		0·8	2·0
LiF	111	2·80	1·63	2·92	0·2	1·8
	200	1·51	1·04	1·48		
	400	1·49	1·19	1·48		
CaF_2	111	1·74	1·57	1·95		
	200	2·45	2·13	2·44	0·4	1·7
	220	6·17	3·17	6·36		
$CaCO_3$	211	3·29	2·39	3·46	0·5	1·6

using neutrons have employed single crystals. In general powdered or polycrystalline samples have been used.

One of the main limitations of the Debye–Scherrer x-ray method is the difficulty encountered by the overlapping of reflections from planes of closely similar spacing, although these planes may have quite different orientations relative to the crystallographic axes. This limitation is met even more forcibly in neutron diffraction for, because of the relative weakness of neutron sources, it is necessary to sacrifice resolution in order to achieve reasonable counting rates. Not only is the collimation poor but also large samples have to be used, with still further detriment to the diffracted beam width. Moreover, as θ increases, the angular width of the diffracted spectra increases even more because of the finite width of the wavelength band in the so-called monochromatic beam. For example, the angular width of the (220) diamond reflection, at an angle of $\theta = 25°$ for a wavelength of about 1 Å, is about 1° in comparison with a width two or three times less than this for an ordinary 19 cm x-ray camera. The restrictions set by the powder method become plain when analyses of crystals of lower symmetry, with many close overlapping lines, are attempted.

A difficulty of rather a different kind arises when crystals containing nuclei particularly hydrogen, which give intense incoherent scattering, are being examined. This is illustrated by some early results given by Shull, Wollan, Morton and Davidson (1948) for sodium hydride, for which the significant coherent peaks in the diffraction pattern are much smaller than the background count, which is greatly accentuated by the 79 barns of incoherent scattering for hydrogen. This difficulty can be greatly reduced by using deuterated samples but, in principle, would be solved most easily by using single crystal techniques. This is a consequence of the isotropic nature of the incoherent scattering, which means that with a powdered specimen the whole of the sample contributes to the background, whereas only a

few correctly oriented crystallites contribute to the diffraction peak. By contrast, for a single crystal the whole of the crystal contributes to the diffraction peak also. This point has been emphasized by Lowde (1951) in a description of a method of making measurements with small single crystals using a non-monochromatized neutron beam, and has been demonstrated in practice by Peterson and Levy (1952a) who found that single crystals of KHF_2 gave very satisfactory results without the necessity of deuteration.

Apart from the isotropic background scattering occasioned by isotope and spin incoherence and the instrumental background occasioned by neutrons randomly scattered by materials in the neighbourhood of the spectrometer, there will also be contributions by inelastic scattering, as described in the previous section, and also by multiple scattering within the sample. This latter effect was first noted and explained for neutrons by Wollan and Shull (1948a). Because of the very low coefficients of true absorption for neutrons it is possible for a significant portion of the diffracted beams to be re-scattered before emerging from the large samples which are used. This re-scattering may take place from any suitably oriented crystal plane and is isotropic.

Form of powder specimens. Two different forms of powdered specimen have been found convenient and both are suitable for making absolute measurements of integrated reflections.

The first method uses a parallel-sided slab of material, either compressed or contained in a thin-walled box made of a low-scattering and low-absorbing material such as aluminium, arranged in the symmetrical transmission position on the spectrometer table. The symmetrical position is maintained by rotating the specimen table at one half the angular velocity of the detecting counter. With this arrangement the number of neutrons P_{hkl} detected per minute is related to the structure factor F of the hkl reflection according to the relation

$$\frac{P_{hkl}}{P_0} = \frac{\lambda^3 L}{4\pi r} \frac{t\rho'}{\rho} \frac{\exp(-\mu t \sec\theta)}{\sin^2 2\theta} j_{hkl} F_{hkl}^2 N_c^2, \qquad \ldots\ldots(38)$$

where P_0 is the total number of neutrons per minute in the incident beam, L is the length of the counter slit, r is the distance of the counter slit from the specimen, t is the specimen thickness, ρ' is the measured density of the powder, ρ is the theoretical density, μ is the linear absorption coefficient, and j_{hkl} is the number of co-operating planes.

In the second method of procedure, which is more convenient if the amount of material is limited, a cylindrical sample either formed into a compressed rod or held in a thin-walled aluminium container is bathed in a neutron beam which is slightly wider than the specimen diameter. Using this alternative geometrical arrangement, the appropriate intensity formula is

$$\frac{P_{hkl}}{I_0} = \frac{\lambda^3 L}{8\pi r} \frac{V\rho'}{\rho} \frac{A_{hkl}}{\sin\theta \sin 2\theta} j_{hkl} F_{hkl}^2 N_c^2, \qquad \ldots\ldots(39)$$

where I_0 is the number of neutrons per minute per unit area in the incident beam, i.e. $I_0 = P_0$/beam area, V is the volume of sample irradiated, and A_{hkl} is an absorption factor. The factor A_{hkl} depends on the quantity μR_s, where μ is the linear absorption coefficient and R_s the specimen radius, and on the Bragg angle θ. A_{hkl} has been calculated by Claassen (1930) and by Bradley (1935) and is tabulated in the *International Tables for the Determination of Crystal Structures* (1935), Vol. II.

For either experimental arrangement the value of the absorption coefficient μ can be measured directly by finding the transmission loss of an incident neutron beam in passing through the specimen, using in the second case a beam much narrower than the specimen width and incident along a diameter.

§4. FUNDAMENTAL EXPERIMENTAL MEASUREMENTS

We shall devote the following paragraphs to a description of the methods which have been used to determine by experiment the scattering amplitudes and cross sections of the elements. These are of course the fundamental quantities which, as we have described in the previous two sections, determine the intensities of our neutron diffraction spectra. Our main purpose will be to measure the sign and amplitude of the coherent nuclear scattering amplitudes: as already shown, our precise knowledge of these is based on experiment. We shall also refer to the determination of absorption coefficients.

4.1. *General Determinations of the Sign of the Coherent Scattering Amplitude* $\bar{a}$

(a) Measurements of refractive index from critical angle.

The primary data on the signs of the scattering amplitudes arise from measurements of the refractive index of materials for neutrons.

It can be shown (e.g. Fermi 1950) that the refractive index n of an element, relative to that of a vacuum, is given by

$$n^2 = 1 - \lambda^2 N\bar{a}/\pi, \qquad \ldots\ldots(40)$$

where N is the number of atoms per cm^3. In the case of a compound $N\bar{a}$ is summed over the various elements present. It is emphasized that it is indeed the *coherent* scattering amplitude which determines the refractive index.

It follows from eqn. (40) that if $\bar{a}$ is positive, corresponding to a phase change of 180° on scattering, which will be found to be the case for the great majority of elements, then the refractive index will be less than unity. In fact $1-n$ is of the order of 10^{-6}. Thus when a monochromatic neutron beam is incident at the surface of a material there will be a critical value θ_c of the glancing angle below which total reflection takes place, and which is given by

$$\theta_c = \lambda(N\bar{a}/\pi)^{1/2}. \qquad \ldots\ldots(41)$$

A measurement of θ_c, which is about 10′ of arc in practice, will therefore determine $\bar{a}$.

The first demonstration of total reflection was attempted by Fermi and Zinn (1946). Intense monochromatic beams were not then available and the Maxwell spectrum of thermal neutrons from a pile was used. Under these conditions no limiting angle would be expected. However, very strong reflection intensities were observed at small angles with mirrors of Be, Cu, Zn, Ni, Fe and C, suggesting that total reflection was taking place and that these elements must therefore possess positive scattering lengths. The measurements were repeated by Fermi and Marshall (1947) using a beam of crystal-reflected monochromatic neutrons of wavelength $\lambda = 1{\cdot}87$ Å and about 2′ in angular width. A sharp drop of intensity in the reflected beam at the critical angle was found, establishing conclusively the positive sign for the above elements.

These measurements form the starting point of the compilation of a list of signs of $\bar{a}$. We shall refer to particular refinements of the mirror reflection technique later.

(*b*) *The relative intensities of Bragg reflections.*

The *relative* signs of the scattering lengths of two elements in a binary compound of known structure can be inferred from the relative intensities of coherent Bragg reflections. Thus for a compound AB with the NaCl-type structure it would be expected that if A, B had $\bar{a}$ values of the same sign then the odd orders of reflection from the (111) planes would be weak in comparison with the intense even orders. On the other hand, opposite signs would entail intense odd orders and weak even orders. Fermi and Marshall (1947) made intensity measurements for various planes of many single crystals and, after correction for geometrical factors, were able to deduce the relative signs from the sequences of intensity in successive orders, without having to make precise correlation between structure factor and intensity. For example, with LiF it was deduced that Li, F scattered with opposite signs, whereas observations with NaCl showed that Na, Cl had the same sign. By intercomparison of various elements it was clear that C, O, Fe, Mg, Ba, Ca, S, Pb, F, N, Na, Cl, K, Br and I all had the same sign. This common sign was then seen to be positive by comparison with the results of mirror reflection for Fe, C.

4.2. *General Determinations of the Magnitude of* $\bar{a}$

A measurement of θ_c in a mirror reflection experiment would enable the magnitude of $\bar{a}$ to be deduced from eqn. (41). However, in general the numerical value can be found more easily by other methods.

(*a*) *Determination of cross sections by transmission measurements.*

We have already referred in §2 to the total integrated scattering by an element. By measuring the reduction of intensity when a monochromatic neutron beam is transmitted through a sample we can determine a total cross section for both scattering and absorption under the given conditions of wavelength and temperature. This effective cross section σ_{eff} is given by

$$I = I_0 \exp(-\sigma_{\text{eff}} N t), \qquad \ldots\ldots (42)$$

where I, I_0 are the beam intensities with and without the sample in position, N is the number of nuclei per cm^3 and t is the sample thickness. Most of the measurements of this type have been made with velocity selectors, which permit easy investigation over a very wide range of neutron wavelength, and have been described in many papers such as those of Havens *et al.* (1948) and of Bendt and Ruderman (1950). They have been reviewed by Goldsmith, Ibser and Feld (1947) and by Adair (1950).

The measured value of σ_{eff} at a particular wavelength will depend not only on the fundamental cross sections σ, S and s for the element but also on crystal diffraction effects, discussed in connection with fig. 1, and on absorption by nuclear capture. However, the results can be interpreted fairly simply at short wavelengths where crystal diffraction effects average out, making the coherent scattering practically isotropic, and where true absorption is negligible. Thus in the region of $\lambda = 0{\cdot}25$ Å, corresponding to a neutron energy a little greater than 1 ev, σ_{eff} becomes equal to the total scattering cross section σ. At much higher neutron energies complications arise from resonance effects.

Elements showing no incoherence. In the simplest case of an element for which there is no isotope or spin incoherence the total scattering cross section

deduced in this way will be identical with $S=4\pi a^2$, the *coherent* cross section. In these cases, therefore, a can be determined directly. However, it must be borne in mind that at these and higher neutron energies the values found for σ, a are those appropriate to *free* atoms and they must be corrected to those of *bound* atoms, which for most elements are the effective ones at the neutron diffraction wavelengths of about 1·1 Å, by multiplication by $(A+1)^2/A^2$ and $(A+1)/A$ for σ and a respectively, where A is the mass number of the nucleus.

In this way absolute values of S, a have been determined for carbon (the effect of ^{13}C in ordinary carbon being negligible, Koehler and Wollan 1952), thorium, ^{58}Ni and other elements. Thus Wollan and Shull (1948 a) report a value of $5{\cdot}2\times10^{-24}$ cm^2 for S for carbon, corresponding to a value for a of $0{\cdot}64\times10^{-12}$ cm.

Elements with spin and isotope incoherence. For elements showing isotope and spin incoherence the value of S can be determined from σ if the disordered scattering s is known. The procedure is illustrated by fig. 8, which shows data

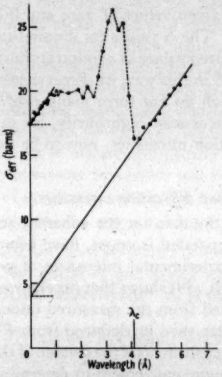


Fig. 8. Variation with wavelength of the measured cross section of nickel, as determined by transmission measurements. (*After* Bendt and Ruderman 1950)

given by Bendt and Ruderman (1950) for nickel. In the short wavelength region the effective cross section is linear with λ and of the form

$$\sigma_{\text{eff}}=17{\cdot}4+0{\cdot}79E^{-1/2}, \qquad \ldots\ldots(43)$$

where E is the neutron energy in electron volts. This expression is to be interpreted as the sum of two quantities, first 17·4 barns, which is the *free scattering* cross section, and secondly the absorption term, which shows the expected proportionality to $1/v$ and corresponds to an absorption cross section σ_a of about 5 barns for neutrons of wavelength 1·8 Å. Thus $\sigma_{\text{free}}=17{\cdot}4$ barns and hence σ, for a *bound* nickel atom, will be

$$\sigma=18{\cdot}0 \text{ barns.} \qquad \ldots\ldots(44)$$

In the long wavelength region, beyond the cut-off wavelength for ordered elastic scattering, where $\lambda > 2d_{max}$, the experimental results can be represented by

$$\sigma_{eff} = 4{\cdot}1 + 0{\cdot}79E^{-1/2}, \qquad \ldots\ldots(45)$$

which corresponds, neglecting inelastic scattering, to a disordered cross section s with the addition of the same absorption term as before. Thus,

$$s = 4{\cdot}1 \text{ barns}; \qquad \ldots\ldots(46)$$

whence, from (44) and (46), we have

$$S = 13{\cdot}9 \text{ barns}. \qquad \ldots\ldots(47)$$

When the absorption of an element is high the accuracy of this method is limited by the long extrapolation back to the axis of ordinates of the portion of the curve beyond the cut-off wavelength, in order to determine the value of s. More precisely, also, the measurements at long wavelength should be done at low temperature in order to reduce the effect of inelastic scattering which will contribute to the apparent value of s. At liquid air temperature the inelastic scattering is negligible. Cassels (1951) and Squires (1952), in some very accurate measurements of this kind, have given values of zero and 0·1 for s for aluminium and magnesium respectively. The S values for aluminium and magnesium can thus be deduced from the measured values of σ found at short wavelengths.

The determination of S in this way, by direct measurement of s, is particularly valuable for elements which do not form many readily available compounds of simple structure and the necessary high purity, for in these cases the method of using the coherent diffraction intensities, now to be described, may not be easily applicable.

(*b*) *Measurements of coherent diffraction intensities.*

The great majority of the data for the coherent scattering amplitudes of the elements, and for many separated isotopes, have been determined by Shull and Wollan (1951) from the experimental intensities of powder diffraction patterns. Reference back to eqns. (38), (39) shows that for elements or compounds of known structure F^2 can be obtained from the measured intensities P_{hkl}. In the case of an element the value of $\bar{a}$ can then be deduced from F^2, and for a compound this can be done for one of the constituents if the $\bar{a}$ value of the other is known.

In practice it is found more convenient to determine the comparative value of the intensities relative to those from a standard substance of known coherent scattering cross section, using eqn. (39), for example, in the modified form

$$P_{hkl} = \text{constant}\ \frac{\rho'}{\rho}\ \frac{A_{hkl}}{\sin\theta \sin 2\theta}\ j_{hkl}\ F_{hkl}^2\ N_c^2, \qquad \ldots\ldots(48)$$

where the instrumental constant is applicable to a powdered specimen of standard size irradiated in a monochromatic beam of standard intensity. Use of this procedure makes it unnecessary to make dimensional measurements of the apparatus or to have knowledge of the intensity distribution across the incident beam.

Shull and Wollan (1951) have given a comprehensive list of $\bar{a}$ values determined in this way from elements and, by intercomparison, from various compounds. Substances used for finding the instrumental constant have included ThO_2, diamond, ^{58}Ni. Deductions from these various substances have been found to be in good agreement. Ordinary nickel, for which S is well established as 13·4 barns, has proved a useful sub-standard on account of its convenient high-intensity diffraction peaks.

4.3. *Other Methods of Determination of Sign and Magnitude of* $\bar{a}$

The experimental methods so far described have been the ones generally used for determining the sign and magnitude of $\bar{a}$. For certain elements other methods have been necessary for various reasons. In addition these methods have given valuable confirmation of the more orthodox techniques when applied to elements of readily determined constants.

(*a*) *Refractive index of gases.*

McReynolds (1951) has described the determination of $\bar{a}$ from the refractive index of gases in the cases of oxygen, nitrogen, helium and argon. This provides the only method for the inert gases and is valuable for nitrogen, whose scattering amplitude was known only inaccurately from diffraction intensity measurements. The method also demonstrates that even for gases the relevant quantity in the refractive index equation (40) is the bound coherent scattering amplitude. This had been suggested from theoretical considerations by Kleinman and Snow (1951). So far as the refractive index is concerned the neutron is to be regarded as interacting with the whole medium rather than with individual atoms, and it is the mean value of $\bar{a}$ which is significant.

In McReynolds' method a beam of 'white' neutrons is reflected from a liquid–gas interface, using a liquid such as ethylene glycol or triethylene glycol which has a refractive index very close to unity on account of the near-cancellation of the carbon and hydrogen scattering amplitudes which are positive and negative respectively. The incident beam, of width 1′, was reflected at an angle of 4′, so that all the incident neutrons of wavelength greater than a certain critical value λ_c are reflected. The experimental procedure was to measure the total reflected intensity I as a function of gas pressure p from 0 to 150 atmospheres. It can be shown that

$$(I/I_0)^{1/2} = 1 - p/p_0 = 1 - N\bar{a}/N_0\bar{a}_0 , \qquad \ldots\ldots(49)$$

where I_0 is the reflected intensity in the absence of the gas : $\bar{a}$, $\bar{a}_0$ are the coherent scattering amplitudes of the nuclei of gas, liquid and N, N_0 are the number of nuclei per cm^3 for gas, liquid respectively. p_0 is the gas pressure at which the reflected intensity is zero.

It follows from eqn. (49) that $I^{1/2}$ is a linear function of p, and extrapolation of the experimental results will give the value p_0 at which I is equal to zero. Then

$$\bar{a} = N_0\bar{a}_0/N_{p_0} . \qquad \ldots\ldots(50)$$

As an absolute method of determination the method is not very accurate since the precise cross section S for the liquid is very sensitive to small errors in the assumed values of S for carbon and hydrogen. However, by comparison with the results for oxygen, assuming $\sigma = 4{\cdot}2$ barns, cross sections for nitrogen, helium and argon were determined as $9{\cdot}7 \pm 1{\cdot}5$, $1{\cdot}1 \pm 0{\cdot}15$ and $0{\cdot}51 \pm 0{\cdot}1$ barn respectively. The experiment demonstrates that helium and argon have positive scattering amplitudes.

(*b*) $\bar{a}$ *for vanadium.*

A procedure very similar to that just described was followed by McReynolds and Weiss (1951) to determine the scattering amplitude of vanadium. On account of the very large spin incoherence, S is smaller by an order of magnitude than that of any other nucleus studied, although the total cross section σ is as large as 5 barns. The coherent scattering is so small as to be difficult to measure by crystal diffraction, and Shull and Wollan (1951) gave simply $S < 0{\cdot}1$ barn. By reflection of a neutron

beam from a vanadium mirror in nitrogen and extrapolation to the gas pressure which would give zero reflectivity McReynolds and Weiss obtained $S=0{\cdot}028\pm0{\cdot}005$ barn and deduced that $\bar{a}$ was positive. However, as will be discussed later, this conflicts with recent measurements of Peterson and Levy (1952 b) by crystal diffraction, which require a negative amplitude. The presence of an oxide film on the vanadium mirror surface may account for the discrepancy.

(*c*) *Coherent scattering of hydrogen.*

A refinement of the mirror reflection technique has been described (Hughes, Burgy and Ringo 1950, Burgy *et al.* 1951) for the precise determination of the coherent neutron–proton scattering amplitude. This is of particular importance in the theoretical treatment of the range of forces in the neutron–proton interaction. Earlier measurements by crystal diffraction (Shull, Wollan, Morton and Davidson 1948) gave $\bar{a}=(-3{\cdot}96\pm0{\cdot}20)\times10^{-13}$ cm and measurements of scattering by ortho- and para-hydrogen (Sutton *et al.* 1947) gave $(-3{\cdot}95\pm0{\cdot}12)\times10^{-13}$ cm. These values, although in agreement, were higher than the theoretical predictions. The former method suffered from a somewhat uncertain correction for thermal vibrations and the latter from a possibility of ortho-contamination in the para-hydrogen. Relative to the crystal diffraction method the mirror reflection technique has the advantage of being independent of any knowledge of the Debye temperature since the scattering angle is practically zero, making correction for thermal displacements negligible.

Two different experimental methods were employed by Burgy, Ringo and Hughes (1951). In the first a 'white' neutron beam was reflected at a glancing angle θ_c of 5·3′ from a surface of triethyl benzene, $C_{12}H_{18}$, and the cut-off wavelength λ_c was found by subsequent reflection from a beryllium mirror. The value of $\bar{a}_H$ for hydrogen is then given in terms of $\bar{a}_C$ for carbon by

$$\theta_c=\lambda_c\left[\frac{N}{\pi}(\bar{a}_C+1{\cdot}5\bar{a}_H)\right]^{1/2}, \qquad \ldots\ldots(51)$$

where N is the number of carbon atoms per cm³ of the liquid.

In the later method the beam was reflected in turn by a series of liquids of varying hydrogen/carbon ratio and measurement made of the glancing angle θ_c which gave some particular constant intensity. It can be shown that θ_c^2 is proportional to the H/C ratio and extrapolation of the experimental results to $\theta_c=0$ therefore gives an H/C ratio equal to $\bar{a}_C/\bar{a}_H$. In this way a value of $\bar{a}_C/\bar{a}_H=-1{\cdot}753\pm0{\cdot}005$ was obtained, leading to $\bar{a}_H=(-3{\cdot}78\pm0{\cdot}02)\times10^{-13}$ cm. The discrepancy between this figure and the latest value of $(-4{\cdot}1\pm0{\cdot}2)\times10^{-13}$ reported by Shull and Wollan (1951) has not been explained.

(*d*) *Scattering amplitudes of deuterium.*

Various measurements have been made of the scattering of neutrons by deuterium, with a view to establishing the two scattering lengths a_1, a_2 corresponding to neutron spin parallel and antiparallel respectively to the deuteron spin. Measurements by crystal diffraction with NaD, heavy ice, ThD_2, LiD (Wollan, Shull and Koehler 1951) and with ND_4Br (Levy and Peterson 1951) give $S=5{\cdot}2\pm0{\cdot}3$ barns and a bound scattering amplitude of $(0{\cdot}64\pm0{\cdot}02)\times10^{-12}$ cm. This result, together with a measurement of total scattering cross section, which for the *bound* atom is $7{\cdot}6\pm0{\cdot}3$ barns (Fermi and Marshall 1949, and Rainwater *et al.* 1948), gives two possible pairs of values for a_1, a_2.

By a different method, in which the angular distribution of neutrons scattered by deuterium gas is measured, Hurst and Alcock (1951) have determined the ratio a_2/a_4, again with two possible values, and Fermi and Marshall (1949) have obtained a further relation between a_4 and a_2 by measuring the transmission of deuterium gas for long wavelength neutrons at 77° K.

The various results have been discussed by both Hurst and Alcock (1951) and Wollan, Shull and Koehler (1951). From them two possible pairs of values of a_2, a_4 can be selected. However, theoretical work by Gordon (1950) and Troesch and Verde (1951) concludes that $a_2 > a_4$. Consequently the values chosen from the above experiments are: a_2 (free) $= (+0{\cdot}83 \pm 0{\cdot}015) \times 10^{-12}$ cm, a_4 (free) $= (+0{\cdot}24 \pm 0{\cdot}015) \times 10^{-12}$ cm.

(*e*) *Measurements of small angle scattering.*

When a narrow beam of thermal neutrons is passed through finely divided material it is found that the beam diverges, the width increasing by some minutes of arc. This is ascribed to refraction and diffraction effects occurring at the individual particles, which are much greater in dimensions than the neutron wavelength. The behaviour and, indeed, the corresponding effect with X-rays, can be described by the Rayleigh (1911) and Gans (1925) theory or by that of von Nardroff (1926) according as diffraction or refraction effects predominate. It has been shown by van de Hulst (1946) that diffraction or refraction predominates according as $\phi \ll 1$ or $\phi \gg 1$, where ϕ is the phase difference in radians between the path through the particle and the same distance in a vacuum.

Weiss (1951) has demonstrated experimentally the two cases for neutrons and has shown that both the differential and total cross sections are correctly given by the theories. From a consideration of the case where refraction is predominant, $\phi \gg 1$, Weiss derives a method of determining the phase and magnitude of the coherent scattering amplitude. Weiss shows that the broadening of the beam is given by

$$(\omega^2 - \omega_0^2)^{1/2} = \text{constant} \,.\, \delta\tau^{1/2}, \qquad \ldots\ldots(52)$$

where ω_0, ω are the angular widths at half maximum intensity of the incident and transmitted beams, $(1-\delta)$ is the refractive index of the particles, τ is the total number of particles traversed. If measurements are made not only by direct transmission through a powdered sample but also with the sample immersed in a liquid of very low absorption such as CS_2, then we have

$$\frac{(\omega^2 - \omega_0^2)^{1/2}}{(\omega^2 - \omega_0^2)^{1/2}_{CS_2}} = \frac{N\bar{a}}{N\bar{a} - N_{CS_2}\bar{a}_{CS_2}} \qquad \ldots\ldots(53)$$

where N, N_{CS_2} are the number of scattering nuclei per cm^3. This follows from the expression given above, at eqn. (40), for the refractive index. Experimentally the measurements were made with neutrons totally reflected at a glass mirror giving a beam of effective wavelength 3·2 Å and about 2′ in angular width.

From eqn. (53) the value of $\bar{a}$ for the particles is thus obtained. The method is not so accurate as the crystal diffraction method but has the advantage of needing very little material, 500 mg being sufficient. Moreover the sign of the scattering amplitude is readily determined since, except for the rare substances with a value of δ less than that of CS_2, it follows that if the beam width increases when CS_2 is introduced then $\bar{a}$ is positive: if the width decreases then $\bar{a}$ is negative. Using this method Weiss showed for the first time that

Si, P, Mo, Pd, Cb, Tc and Ru have positive scattering amplitudes and confirmed the signs for many other elements, including the negative value for manganese. The magnitude of $\bar{a}$ was found for several elements, including the first reported value of silicon of $(0{\cdot}4 \pm 0{\cdot}02) \times 10^{-12}$ cm. Preferably a liquid should be chosen with a refractive index close to that of the material being investigated. Only a few liquids, such as CS_2 and D_2O, have a sufficiently low absorption to be used satisfactorily.

For Weiss' detailed experimental verification of the diffraction and refraction aspects of the small-angle scattering theory the reader is referred to the original paper (1951).

(*f*) *Measurements of metallic solid solutions.*

Bendt and Ruderman (1950) have described a method of finding the relative scattering phases of elements by measurements of cross section for solid solutions. Referring back to the discussion of fig. 8 we recall that S for nickel was determined by subtraction of s from σ (eqns. (44), (46), (47)). If a similar (σ, λ) curve is obtained for an alloy we can determine the effective value of σ. This will be given by

$$\sigma_{\text{eff}} = 4\pi[p_1\bar{a}_1 + p_2\bar{a}_2]^2, \qquad \ldots\ldots(54)$$

where p_1, p_2 and $\bar{a}_1$, $\bar{a}_2$ are the atomic proportions and scattering amplitudes (signs assumed to be unknown) of the two constituents. By comparison of the experimentally determined value of σ_{eff} for the alloy with eqn. (54) it will be clear whether $\bar{a}_1$, $\bar{a}_2$ have the same or opposite sign. The negative scattering amplitude of manganese has been confirmed by this method also, using a manganese nickel alloy.

(*g*) *Total cross section measurements in the interference region.*

We have shown in our discussion of total coherent scattering cross section $E(S)$ that beyond a certain cut-off wavelength, such as is indicated at λ_c in fig. 1, the coherent elastic scattering falls to zero. With reduction of neutron wavelength below λ_c there is a discontinuous jump in the cross section to a value

$$E(S) = \tfrac{1}{2}N_c\lambda^2 F_{hkl}{}^2 d_{hkl}. \qquad \ldots\ldots(55)$$

The magnitude of this jump is therefore determined by the value of $F_{hkl}{}^2$ for the planes of maximum spacing. In the case of a binary compound the value of F may be quite different, depending on whether the two elements have the same or opposite signs for $\bar{a}$. By comparing the shape of the experimental $E(S)$ curve in this region with the results of calculation Fermi, Sturm and Sachs (1947) showed that Be and O must have the same phase. Winsberg, Meneghetti and Sidhu (1949) confirmed the negative scattering amplitude of titanium in a similar way and also deduced the magnitude of $\bar{a}$ for a number of elements.

Weiss (1952) has investigated the effect of extinction on the shape and magnitude of the peaks in the $E(S)$ curve. Since these occur when, for a particular set of planes, the value of $\lambda/2d$ falls below unity, it follows that $\theta = 90°$ at a peak. Weiss shows that the diffraction width of the reflection from a mosaic block increases rapidly as $\theta \to 90°$ and becomes much greater than the spread of the mosaic blocks. Under these conditions the mosaic spread ceases to be the factor which determines the extinction effects, and it is the grain size which is of main importance. From a detailed study of the way in which the peak near the cut-off wavelength is reduced and rounded off the mosaic block size and spread can be determined.

(*h*) *Diffuse scattering in powder diffraction patterns.*

A measurement of incoherent scattering from a powder diffraction pattern, rather than by transmission methods, has been reported by Shull and Wollan (1951) for vanadium carbide. As already mentioned, the spin incoherent scattering of vanadium is very large and contributes a large background to the powder diffraction pattern. After correction, by calculation, for thermal scattering and multiple scattering the remainder of the scattering was expressed as a differential scattering cross section ds per unit solid angle. It was found that the scattering was isotropic, with ds equal to 0·4 barn, corresponding to a total incoherent cross section s of 5·0 barns, in agreement with the known σ value of 5·0 barns and the S value (Peterson and Levy 1952 b) of only 0·029 barn.

4.4. *Coefficients of True Absorption*

We listed in table 2 the cross sections of the elements for true absorption, not including scattering, as recorded in the volume *Nuclear Data* (1950). The experimental measurement of absorption is usually made by the ' pile oscillator ' method which has been described by a number of workers (Weinberg and Schweinler 1948, Hoover *et al.* 1948, and Pomerance 1951). A sample is oscillated back and forth through a neutron-detecting chamber placed in the graphite reflector of a pile. The variation of detector current depends on the capture cross section of the sample and is relatively insensitive to the effect of absorption by scattering. In practice the apparatus is calibrated using gold, whose capture cross section is known accurately by other direct methods (Brill and Lichtenberger 1947, McDaniel *et al.* 1947). Pomerance (1951) has discussed the accuracy of the absorption coefficients determined by this and other methods. The values given in table 2 have been adjusted to a neutron wavelength of 1·08 Å by using the $1/v$ law for σ_a.

§ 5. THE ROLE OF NEUTRON DIFFRACTION IN CRYSTALLOGRAPHIC ANALYSIS AND IN A STUDY OF THE PHYSICS OF THE SOLID STATE

X-ray diffraction methods have achieved outstanding triumphs in the analysis of the structure of molecules and of crystals of considerable complexity, but there are four main limitations on these methods which as yet have not been overcome:

(i) It is not possible to resolve light atoms in the presence of a high proportion of heavy atoms. In particular the positions of hydrogen atoms even in hydrocarbons can only be determined by an accurate Fourier difference synthesis, and in the heavy atom hydrides they cannot be located by x-rays at all.

(ii) It is practically impossible to distinguish atoms, such as Fe and Co, which have neighbouring atomic numbers and therefore have closely similar scattering power. The difference between an ordered and a disordered alloy of such metals cannot be detected by x-rays.

(iii) The positions of nuclei cannot be located by x-rays except on the assumption that the nucleus is central to the scattering cloud of electron density. It is not always possible to be sure that apparent deviation from spherical scattering symmetry in the atom may not, in fact, be due to a displacement of the whole atom and a lowering of symmetry of the crystal structure.

(iv) X-rays cannot detect magnetic spin moments or alignment of magnetic spins and can only observe magnetization, if at all, as a second-order distortion of the crystalline lattice.

To this one may add the fact that x-rays cannot distinguish isotopes and could therefore not be used to determine the proportion of isotopes in any crystal except by the second-order method of variation of lattice parameter.

All these fields have proved suitable for neutron diffraction investigation. It follows that neutron diffraction can be used to supplement the information available from x-ray methods. In addition to the above there are other important differences in the two techniques which restrict or enhance the usefulness of one or the other.

(1) In general, neutron sources are relatively weak, and therefore it is not yet possible to use very small single crystals. With x-rays crystals need not be more than a small fraction of a millimetre in size.

(2) Neutrons cannot be rigorously monochromatized without a loss of intensity too great for diffraction methods to be used. There is always a spectral band rather than a spectral line even after monochromatization. This results in a broadening of high-angle reflections, a reduction of peak-to-background ratio and a loss of precision in intensity measurements.

(3) Methods of intensity measurement and record of neutron beams are limited to the use of secondary particles. Photographic techniques, which are of the greatest use in x-ray investigations, are relatively undeveloped for neutrons.

(4) The need for remote control and automatic recording imposes a limitation on experimental manipulation in the case of neutron diffraction methods. Adjustment of crystal specimens by use of a fluorescent screen, for example, is quite impossible.

(5) Neutron absorption is variable from element to element (table 2) and in some cases inhibits the use of single crystal methods.

(6) Coherent scattering cross sections have to be determined empirically in the case of nuclear interaction with neutrons, and are known only approximately for some elements and not at all for others.

(7) In the case of nuclear scattering effects the intensity series are not usually sufficiently extended or convergent for Fourier analysis to be undertaken with any certainty of obtaining unambiguous results.

(8) Since the range of values of coherent neutron scattering amplitudes is so limited, direct methods of Fourier analysis, comparable with the 'heavy atom' method for x-rays, would in any case be less useful.

(9) Except in the special case of magnetic atoms, electron density distribution cannot be determined by neutron diffraction methods, nor can there be any difference in scattering by atoms or by ions of different valencies. Problems in these fields must be solved by x-ray methods.

(10) Where a crystal contains an element having nuclear spin or a mixture of isotopes there is a different neutron scattering cross section for each spin state and for each isotope. The crystal therefore behaves towards neutrons as a 'mixed crystal', and the incoherent background is correspondingly high. This is particularly troublesome in the case of hydrogenous compounds unless single crystal methods can be used.

(11) Owing to the (in general) low absorption and the comparatively large size of the specimens used in neutron diffraction experiments multiple scattering of

neutrons within the sample is liable to contribute heavily to the background. On the other hand, the low absorption coefficients for neutrons permit interior, rather than merely surface, examination of large specimens of metals, etc.

(12) The possibility of negative amplitudes for the neutron scattering from certain nuclei introduces a complexity into calculation, although in cases where a difference in amplitudes is important, as for example in the case of order–disorder investigations, it may be advantageous in extending the range of amplitudes. Where isotopes exist having nuclear scattering amplitudes of opposite signs, it should theoretically be possible so to adjust their proportions that these atoms are effectively eliminated from contributing to the coherent scattering pattern.

(13) The wide variation of neutron scattering amplitudes for isotopes of certain elements allows the possibility of using such isotopes in the isomorphous replacement method, for example, H and D, Fe and ^{54}Fe (R. Pepinsky, private communication).

Most diffraction studies using neutrons have been made by the method analogous to that invariably used in x-ray investigations, that is, by the measurement of intensity of coherent diffracted beams and of the incoherent diffuse background as a function of diffraction angle, the wavelength being kept constant. Laue photography with neutrons, although demonstrated (Shull and Wollan 1948, Wollan, Shull and Marney 1948) has not been used for structure studies.

A second method, however, is possible with neutrons that is practically never used in x-ray investigations, namely, to measure the intensity loss in the direct beam on transmission through the scattering material, when the neutron energy in the incident beam is varied as described in §4 (Halpern, Hamermesh and Johnson 1941, Fermi, Sturm and Sachs 1947). This method was used to determine the first absolute value of the coherent scattering cross section of carbon in diamond.

One method of obtaining an approximately monochromatic beam of long wavelength neutrons is to filter through a considerable thickness of some non-absorbing material such as BeO, free from spin or isotopes.

For thin specimens and for values of λ less than $2d_{max}$ the 'transmission versus incident energy' curves show peak values as one after another of the crystal planes takes part in the scattering, and the spectra obtained as in fig. 1 are therefore characteristic of the structure of the specimen. Experiments have been made using CaO and TiC samples (Winsberg, Meneghetti and Sidhu 1949), but the absolute values of the coherent cross section obtained by the use of this method do not agree very well with those obtained by the usual diffraction methods.

The transmission does, of course, vary with preferred orientation in the specimen (Arnold, Myers and Weber 1949) and may be used to study the effects of cold-working, for example, in metal sheet.

It follows also that the transmission is dependent on the physical state of the specimen and will vary, for example, with solid solution as compared with mechanical mixture (Sidhu 1948). A disc of transparent vitreous silica examined by the neutron transmission method (Ruderman 1950) showed well-defined diffraction peaks at interplanar spacings of 0·7, 1·1, 1·7, 1·9, 2·2, 2·7 Å and some others not so well resolved, although x-ray photographs of samples cut from the same specimen showed only the usual diffuse pattern.

Generally speaking, however, this method has not proved useful for ordinary structure determination.

§6. CRYSTAL STRUCTURE DETERMINATIONS INVOLVING LIGHT ELEMENTS

Many of the first studies of simple structures by means of neutron diffraction methods were primarily intended for the determination of neutron scattering amplitudes. The structures had already been determined by x-ray methods and could be assumed, as for example in the case of elements of simple structure, alkali halides and many other simple inorganic compounds. X-ray diffraction methods had shown that the metallic atoms in LiH, NaH, KH, CsH, etc., have a face-centred cubic arrangement ; but it is only in the case of LiH that the H atoms could also be located by x-ray methods. Shull, Wollan, Morton and Davidson (1948) examined powder specimens of NaH and NaD using neutrons monochromatized by reflection from a rocksalt wedge to give a narrow band at $\lambda = 1{\cdot}057$ Å, and obtained diffraction curves which were quite dissimilar in appearance, certain intensity ratios being reversed. The four possibilities to be considered were the rocksalt and the zinc-blende types of structure, each with Na and H (or D) scattering with amplitudes of like or of unlike sign. Trial and error methods, making allowance for the thermal vibration of the atoms, showed that only the rocksalt type of structure fitted both substances, but that Na and D scattered with like, Na and H with unlike, phases. In addition, the diffuse scattering was six times as high for NaH as for NaD because of spin incoherence and multiple scattering. Measurements of the transmission cross section (total scattering and capture cross section) give a value per NaH molecule of 32 barns, of which about 31 barns is in the diffuse part of the pattern.

6.1. *Ice*

In view of the difficulties arising from the spin-incoherence of H, deuterated compounds have largely been used in subsequent studies, except where it has proved possible to use single crystals. The neutron diffraction investigation of heavy ice (D_2O) was carried out on powdered samples at $-90°$C and the intensities of reflection from fourteen planes, giving six peaks, have been compared with those calculated from four possible structures (Wollan, Davidson and Shull 1949). The positions of the oxygen atoms are already known from x-ray measurements, and the question to be decided is the location of the deuterium atoms.

Absolute intensities were obtained by comparing the pattern for a known amount of heavy ice with the pattern given by a known amount of diamond, which is used as standard because the carbon coherent cross section is accurately known. The four structures considered were : (i) the Barnes (1929) model, in which the deuterium atoms would be midway between oxygen atoms; (ii) a Bernal–Fowler (1933) model, in which molecules persist with the O–D distance of 0·96 Å, and in which twelve molecules are symmetrically orientated to give hexagonal symmetry (this requires a larger unit cell and would involve a super-lattice) ; (iii) a model in which the molecule rotates; (iv) a model in which there is $\frac{1}{2}$D in all positions at 0·96 Å from oxygen atoms along O–O bonds. The last model corresponds to a disordered structure in which the orientation of the molecules is random, subject only to the proviso that the D nuclei lie along the O–O bond. Such a structure could be either static, as Bernal and Fowler (1933) suggested, or it could be dynamic, with an interchange of D positions, as suggested by Pauling (1935).

The integrated intensities, on an absolute scale, of the six observed peaks were compared with intensities calculated, using 5·2 and 4·2 barns for the coherent scattering cross sections of deuterium and oxygen respectively, for these four

structures, the reflections being grouped together to correspond, where necessary, with unresolved peaks.

The best agreement was found for the random model (iv), and this was taken as strong evidence for the correctness of the Pauling structure. Such trial and error methods do not, unfortunately, eliminate the possibility that other structures, not tried, may also give good agreement with observations. It is possible, for instance, that the oxygen atoms may form a strictly hexagonal structure, but that the deuterium nuclei, though not disordered, may have an arrangement of lower symmetry so that the structure as a whole is only pseudo-hexagonal, although strictly hexagonal symmetry may be simulated by twinning. It is very desirable that, if possible, the experimental measurements should be repeated with the improved resolution now available, so that a more rigorous test of the random model may be made ; the magnitude of the diffuse scattering does, however, constitute an additional indication of randomness. The coherent scattering cross section of deuterium in a crystal with no randomness would be $4\pi\bar{a}_D^2$. With an average coherent scattering amplitude of $\bar{a}_D/2$ in $2N$ sites the average coherent scattering cross section per deuterium nucleus would be $[4\pi\Sigma_{2N}(\bar{a}_D/2)^2]/N = \frac{1}{2}4\pi\bar{a}_D^2$. The difference, $\frac{1}{2}4\pi\bar{a}_D^2 = 2{\cdot}6$ barns, will appear as diffuse scattering additional to the background scattering resulting from spin incoherence (2·3 barns) and thermal incoherence (negligible at small angles); and in fact there is a total incoherent scattering of about 4·9 barns after correction for multiple scattering has been made. The neutron diffraction measurements do not distinguish between static and dynamic disorder.

6.2. *Hydrides of Heavy Elements*

The hydrides of some of the heavy elements have proved to be surprisingly complex. X-ray methods sufficed to determine the heavy atom positions, but the positions of the hydrogen atoms had only been previously guessed at.

Uranium hydride UH_3 (Rundle 1951) has a primitive cubic lattice, $a_0 = 6{\cdot}64$ Å, and an arrangement of U atoms at $0, 0, 0$; $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$; $\pm|\frac{1}{4}, 0, \frac{1}{2}|$; $\pm|\frac{1}{2}, \frac{1}{4}, 0|$; $\pm|0, \frac{1}{2}, \frac{1}{4}|$. It was previously suggested that the hydrogen atoms lay at the mid-points of the U–U bonds, an arrangement which would give U–H distances, in excellent agreement with those calculated from Pauling's rule and metallic radii (Pauling and Ewing 1948). Because of the higher coherent scattering factor and lower spin diffuse scattering of deuterium for neutrons, powdered UD_3 rather than UH_3 was used for neutron diffraction study, the data being obtained by C. G. Shull. Since the coherent scattering amplitude of U was not known it had first to be determined by a study of UC and UO_2, both of known structure. The known U positions are possible in five cubic space groups, and there are five possible sets of equivalent positions of the 24 D atoms, three general and two special. Those, and some combinations of 12-, 16- and 8-fold positions, were tested, though not exhaustively, and certain arrangements were eliminated completely. The bridge structure previously suggested, or any structure approaching it, was found to be quite impossible, and the structure which gave the best agreement with observation was one in which each D atom was at the centre of a deformed tetrahedron of four U neighbours, the U–D distance being 2·32 Å. This is abnormally high, and since the hydride is not saline but metallic, the uranium–deuterium interactions must be of an unusual kind. Each U atom is surrounded by 12 deuterium neighbours, and there are no metallic U–U bonds left in the structure.

The three isomorphous hydrides ThH_2, $Th D_2$ and ZrD_2 have been examined by Rundle, Shull and Wollan (1952), using x-ray and neutron diffraction powder methods. They are body-centred tetragonal, and the x-ray powder diagrams have given the lattice constants of ThH_2 as $a_0 = 4{\cdot}10$, $c_0 = 5{\cdot}03$; $Z = 2$, $\rho = 9{\cdot}20$ g cm^3. The corresponding constants for ZrH_2 are $a_0 = 3{\cdot}520$, $c_0 = 4{\cdot}449$. The metal atoms are at 0, 0, 0 and $\frac{1}{2}$, $\frac{1}{2}$, $\frac{1}{2}$. Five possible sets of positions of the hydrogen atoms are as follows:

$$(0, 0, 0;\ \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}) + (0, \tfrac{1}{2}, \tfrac{1}{4};\ \tfrac{1}{2}, 0, \tfrac{1}{4}) \quad \ldots\ldots(1)$$

$$(0, 0, 0;\ \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}) + (0, \tfrac{1}{2}, 0;\ \tfrac{1}{2}, 0, 0) \quad \ldots\ldots(2)$$

$$(0, 0, 0;\ \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}) \pm 0, 0, z \quad \ldots\ldots(3)$$

$$(0, 0, 0;\ \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}) + (0, \tfrac{1}{2}, z;\ \tfrac{1}{2}, 0, z) \quad \ldots\ldots(4)$$

$$(0, 0, 0;\ \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}) + (0, \tfrac{1}{2}, z;\ \tfrac{1}{2}, 0, \bar{z}) \quad \ldots\ldots(5)$$

Only the first of these (or the structure (4) with $z = \frac{1}{4}$, which comes to the same thing) gives satisfactory agreement with experiment, and since there is here no parameter involved the structure gives an improved neutron scattering factor for deuterium. The coherent scattering factors for Zr and Th were not previously known and had to be determined from ZrN, ZrO_2, Th and ThO_2. The arrangements are all distorted fluoride structures, and again the metal–deuterium distances are very long, being 2·41 Å in ThD_2 and 2·09 Å in ZrD_2, and the hydrides and deuterides are black and metallic and not saline.

6.3. *Heavy Element Carbides*

When atoms as heavy as Th are involved, x-ray diffraction methods fail to locate even carbon atoms with certainty. Hunt and Rundle (1951) and Baenziger (quoted by them) have shown by good x-ray powder and single crystal patterns that ThC_2 is monoclinic (pseudo-tetragonal) with $a_0 = 6{\cdot}53$, $b_0 = 4{\cdot}24$, $c_0 = 6{\cdot}56$ Å, $\beta = 104°$, $Z = 4$. The Th positions were found by Patterson projections, but no evidence of C appeared at all. Neutron-diffraction powder diagrams supplied by Shull gave no additional diffraction peaks, showing that the C positions did not require a larger unit cell than that found. The x-ray data did not differentiate between the space groups $C2/c$ and Cc, but the latter would involve two sets of carbon positions with six unknown parameters. It was found that the neutron diffraction observations could be satisfactorily accounted for by a $C2/c$ structure containing definite C_2 groups, but these involve a C–C distance of 1·5 Å, which is the single-bond distance and not the triple-bond distance previously assumed. The Th–C distances are close to the sum of the Pauling covalent single-bond radii for Th and C (2·42 Å), although the compound itself certainly possesses some ionic character. The authors point out that the structure suggested is metallic in character and that in this connection it is interesting that pressed powders of ThC_2 conduct as well as pressed Th metal powders (P. Chiotti, private communication).

6.4. *KHF_2, Potassium Bifluoride*

X-ray diffraction methods have shown that the F–F distance in potassium bifluoride is only 2·26 Å, which indicates the existence of a strong H bond, but it has not been known whether the H nucleus is half-way between the fluorine atoms or at random in two possible positions close to one or the other fluorine

atom. Physical methods of deciding this point at first gave contradictory results, but recently these methods have pointed rather decisively to the first alternative. Neutron diffraction studies using single crystals (Peterson and Levy 1951, 1952 a) have confirmed this. An attempt was made to use the powder method with a deuterated specimen, but it was found that the powder specimens used were always preferentially orientated, so that the intensities observed were not suitable for comparison with calculation. It was not necessary to use a deuterated single crystal because the ratio of diffuse scattering to Bragg intensity from a single crystal is much less than from a powdered specimen. For the single crystal method the spectrometer was used with minimum collimation, the wavelength spread being of the order of $\pm 1{\cdot}5\%$ at half-maximum intensity, and the sensitive angle subtended by the detector about 2°. Integrated intensities which were reproducible to a few per cent were made with the crystal and detector rotating synchronously through the Bragg angle with a crystal angular speed of 8° per hour. The linear absorption coefficient was measured on a sample of KHF_2 powder. Planes appreciably affected by extinction were not used for parameter determination.

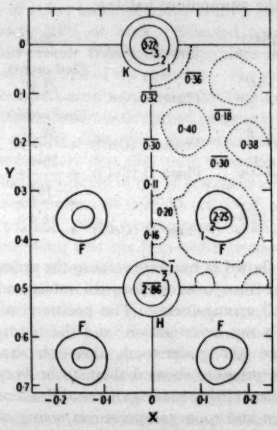


Fig. 9. A projection on (001) of the neutron scattering density of KHF_2. The numbers indicate the values of the function at the contours and the extremes in units of 10^{-13} cm/Å^3. The dashed lines are zero contours. (*After* Peterson and Levy 1952 a)

The tetragonal symmetry was checked by comparison of neutron intensities from (*hk*0) and (*kh*0) planes, in order to eliminate the possibility that the true unit cell, taking the H positions into account, might have lower symmetry. Using the fluorine parameter value given by x-ray studies and a single Debye temperature factor it was not possible to find a hydrogen parameter that would give satisfactory agreement with experimental intensities, but agreement was found when separate temperature factors for K, H and F were introduced and systematically varied, that for F being asymmetric. Final values used were

$B_K = 1{\cdot}25$, $B_H = 3{\cdot}0$, $B_F = 1{\cdot}85$, $B'_F = 1{\cdot}0$, where B_F represents the effect of ordinary spherically symmetrical temperature motion and B'_F the effect of an asymmetric component in the c direction. The rather large Debye factor for H, which is probably mainly zero-point disorder, reduces the precision with which the H parameter can be determined by reducing the H contribution to the high-angle reflections. It may, however, be said that the H position is within 0·1 Å of the centre of the F–H–F bond.

Altogether some 44 reflections were measured in five different zones, and the reflections in the $\{hk0\}$ zone were used to construct the Fourier projection which is shown in fig. 9. No convergence factor was applied to the data, and the effects of termination of series are apparent as diffraction rings around the K and F nuclei. The H nucleus, because of its negative scattering amplitude, appears as a negative peak in the diagram. This is the first Fourier diagram to be constructed from neutron intensity data.

6.5. *Ammonium Halides*

The following transformations (for references see Levy and Peterson 1953) are known to occur in the ammonium halides:

NH_4Cl : Phase I (NaCl)	$\overset{184{\cdot}3^\circ}{\longleftrightarrow}$	Phase II (CsCl)	$\overset{-30{\cdot}5^\circ}{\longleftrightarrow}$ (2nd order)	Phase III (CsCl)
ND_4Cl : Phase I (NaCl)	$\overset{175^\circ}{\longleftrightarrow}$	Phase II (CsCl)	$\overset{-23{\cdot}8^\circ}{\longleftrightarrow}$ (2nd order)	Phase III (CsCl)
NH_4Br : Phase I (NaCl)	$\overset{137{\cdot}8^\circ}{\longleftrightarrow}$	Phase II (CsCl)	$\overset{-38{\cdot}1^\circ}{\longleftrightarrow}$	Phase III (tetragonal)
ND_4Br : Phase I (NaCl)	$\overset{125^\circ}{\longleftrightarrow}$	Phase II (CsCl)	$\overset{-58{\cdot}4^\circ}{\longleftrightarrow}$	Phase III (tetragonal)
			$\overset{-104^\circ}{\longleftrightarrow}$	Phase IV (CsCl)
NH_4I : Phase I (NaCl)	$\overset{-17{\cdot}6^\circ}{\longleftrightarrow}$	Phase II (CsCl)	$\overset{-41{\cdot}6^\circ}{\longleftrightarrow}$	Phase III (tetragonal)

The structure types shown in brackets refer to the arrangements of the N and halide atoms as deduced from x-ray studies, the tetragonal phase III being also a slightly distorted CsCl arrangement. The positions of the H (or D) atoms could not be found by x-ray investigation, and the nature of the second-order transitions was therefore not understood, although sharp transitions in the specific heat at constant pressure showed them to be very real.

Goldschmidt and Hurst (1951) made neutron diffraction studies of powdered ND_4Cl at liquid nitrogen and room temperatures, using neutrons of wavelength 1·07 Å monochromatized by transmission through an NaCl crystal. Their preliminary comparisons of observed and calculated intensities in the low-temperature phase cast doubt on the values available for the coherent cross sections of Cl and N, and these were therefore redetermined from liquid air powder diffraction patterns of NaCl and ZrN, special care being taken to obtain really fine-grained powders. The value for N was found to be higher than that previously determined. Good agreement was found in the low-temperature phase with a tetrahedral ammonium group having a N–D separation of $1{\cdot}03 \pm 0{\cdot}02$ Å, as compared with $1{\cdot}025 \pm 0{\cdot}01$ Å obtained from magnetic resonance measurements on NH_4Cl. For the structure at room temperature Goldschmidt and Hurst considered four models, one with the ammonium ion rotating freely, one with total disorder in the

arrangement of the tetrahedra relative to the diagonals of the cube, one with a larger unit cell than the low-temperature phase, and one identical in unit cell and space group with the low-temperature phase. The calculations were found to lead to the last arrangement as being the most probable, free rotation and disorder being ruled out, although an anisotropic vibration of the deuteron was indicated, pointing to a rotatory oscillation of the ammonium ions.

Levy and Peterson (1952) disagree with these results. They repeated Goldschmidt and Hurst's powder diffraction measurements, using a neutron beam monochromatized by reflection from a (111) face of a copper crystal to give a wavelength of 1·16 Å. Their intensity measurements agreed with those of Goldschmidt and Hurst at low temperatures, but showed significant differences at room temperatures especially in the 111 reflection, which was crucial to the argument favouring an ordered structure. Levy and Peterson found that they could obtain no fit with an ordered model and that their best agreement was obtained with a disordered arrangement. The diffuse background was also observed to be angle-dependent, with a hump in the neighbourhood of 40° scattering angle, and this also implies the existence of disorder (Zachariasen 1945).

A further independent check was carried out by Levy and Peterson, using NH_4Cl single crystals of cubic habit obtained from a saturated aqueous solution containing approximately 10% of urea. Selected specimens were ground to cylindrical shape so that precise absorption corrections could be made with [100] and [110] zone axes along the axis of the cylinder. No evidence of extinction was found although the specimens used were somewhat larger than would be calculated as the maximum permissible for a mosaic spread parameter of 5 minutes. The data agreed well with the powder diffraction measurements obtained by the same authors, and established clearly that the room temperature structure of NH_4Cl involves a disorder in the orientation of the ammonium ions. The N–H and N–D distances were both found to be $1{\cdot}03 \pm 0{\cdot}02$ Å (cf. 1·025 and 1·039 Å from nuclear magnetic resonance studies). ND_4Cl patterns at -180°c and -78°c showed no essential differences, implying that the line-width transition at about -140°c in NH_4Cl observed in nuclear magnetic resonance experiments (Gutowsky and Pake 1948) involves no appreciable change in atomic positions. The Debye factors, which were systematically varied to give the best agreement, were found to be $B_N = 2{\cdot}0$, $B_C = 2{\cdot}0$, $B_D = 2{\cdot}25$, $B_H = 2{\cdot}5$, and the ammonium ions are in rotatory oscillation with characteristic half-angles of $10° \pm 3°$ for ND_4 and $12° \pm 3°$ for NH_4. A by-product of the work was evidence that the scattering amplitude of 0·38 for H as found by Hughes, Burgy and Ringo (1950) was found to be preferable to the earlier value of 0·40, and that the scattering amplitude of N had to be raised to 0·94. This was subsequently confirmed by measurements on VN and KN_3 (Peterson and Levy 1952 b). Goldschmidt and Hurst (1952) have since accepted Levy and Peterson's interpretation of the experimental results.

A preliminary investigation of the second-order transition of ND_4Br (Levy and Peterson 1951) has now been followed by a full determination of the crystal structure of ND_4Br in four phases (Levy and Peterson 1953). ND_4Br was chosen both because of its suitability for powder diffraction studies and because the four phases are obtainable. It is interesting that the isotopic purity of the sample was tested by neutron transmission as well as by determination of the density of a recovered final portion of solvent ; the two methods were in excellent agreement, giving 97·5% D.

The unit cell given at liquid nitrogen temperatures (phase IV) $a_0 = 4{\cdot}00$ Å, was consistent with that (4·010) given by x-ray methods. The space group is $P\bar{4}3m$ with N at 0, 0, 0; Br at $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$ and D at x, x, x; $x, \bar{x}, \bar{x}$; $\bar{x}, x, \bar{x}$; $\bar{x}, \bar{x}, x$. Using separate Debye temperatures factors $B_N = 0{\cdot}8$, $B_{Br} = 0{\cdot}2$ and $B_D = 1{\cdot}0$, found by trial, excellent agreement was attained with $x = 0{\cdot}148$, giving a N–D distance of $1{\cdot}026 \pm 0{\cdot}02$ Å. An oscillatory movement of the ammonium ion with half-angle α of 5° is consistent with, but not necessarily required by, the data.

The structure at room temperature (phase II) was next examined and again it was found that the cubic unit cell with $a_0 = 4{\cdot}060$ Å, found by x-rays, was consistent with the neutron diffraction data. A marked hump in the diffuse background, similar to that observed for ND_4Cl, was taken to indicate a disordered structure. No agreement could be obtained with a model as in phase IV or with other x values. Models containing freely rotating and orientationally disordered ammonium ions were next tried. The latter gave the best agreement, but a severe temperature motion corresponding to a rotatory oscillation of the ammonium ion had to be assumed. This was approximated by considering that each deuterium ion is distributed not only in a spherical cloud (given by the usual temperature factor) but also that the centre of this cloud is distributed about a circle centred on one of the body diagonals of the unit cell, whose radius subtends an angle of $12{\cdot}5° \pm 3°$ at the N nucleus. The final parameters were $x = 0{\cdot}1475 \pm 0{\cdot}003$, $B_N = 1{\cdot}7$, $B_{Br} = 1{\cdot}7$, $B_D = 2{\cdot}5$, the N–D distance being $1{\cdot}037 \pm 0{\cdot}02$ Å.

Phase III of ND_4Br is tetragonal, $c_0 = 4{\cdot}130$ at -124°C, $a_0/c_0 = 1{\cdot}409$ (cf. $\sqrt{2} = 1{\cdot}414$), with N at 0, 0, 0; $\frac{1}{2}, \frac{1}{2}, 0$ and Br at $0, \frac{1}{2}, u$ & $\frac{1}{2}, 0, \bar{u}$; $u = 0{\cdot}52$ (or 0·48) from x-ray studies (Weigle and Saini 1936). This corresponds to a slightly distorted CsCl structure, the actual space group being $P4/nmm$. The D atoms would be at $0, x, z$; $x, 0, \bar{z}$; $0, \bar{x}, z$; $\bar{x}, 0, \bar{z}$; $\frac{1}{2}, \frac{1}{2}+x, \bar{z}$; $\frac{1}{2}+x, \frac{1}{2}, z$; $\frac{1}{2}, \frac{1}{2}-x, \bar{z}$; $\frac{1}{2}-x, \frac{1}{2}, z$. If the ammonium ion is tetrahedral, x and z must be nearly equal. Neutron diffraction data were obtained at -78°C and gave no peaks inconsistent with the x-ray assignment of unit cell and space group, but limited resolution restricted the usefulness of the data and a complete, systematic variation of all parameters was not attempted. The N–D distance was assumed to be 1·03 Å and a distribution corresponding to a rotatory oscillation of ND_4^+ had to be assumed for the D atoms to give satisfactory agreement with the experimental data. Values of u greater than $\frac{1}{2}$ were found to be quite unsatisfactory, and the best agreement was obtained for $x = z = 0{\cdot}147$; $u = 0{\cdot}47$, $B_N = 1{\cdot}2$, $B_{Br} = 0{\cdot}4$, $B_N = 3{\cdot}7$.

A limited number of other trial structures were investigated, but none gave any indication of possible agreement.

In the final structure the Br atom is not tetrahedrally surrounded by D atoms but is at the apex of a square pyramid, the Br atoms being displaced towards the D atoms as compared with the arrangement in an ideal CsCl structure.

Phase I is the high-temperature form of the rock-salt type, with lattice constant 6·90 Å and space group $Fm3m$, according to previous x-ray data. Neutron diffraction studies at 150°C and 200°C were consistent with this unit cell and space group, but marked convergence due to a high temperature factor limited the usefulness of the results. Various models having tetrahedral ammonium ions and N–D equal to 1·03 Å were tested. Three gave reasonably satisfactory agreement including one in which the ND_4^+ ions rotate about a threefold axis set at random parallel to the six crystal ⟨001⟩ directions, with $B_N = 3{\cdot}0$, $B_{Br} = 3{\cdot}0$, $B_D = 4{\cdot}0$. The relatively large value of B_{Br} may well partly represent random displacements of

the ion from ideal positions, resulting from its random environment. Figure 10 shows the four phases, but in the case of phase I only one of the six random orientations of the rotator in one of the three possible models is shown. It is interesting that even at this temperature there is not free rotation of the ammonium ions.

6.6. *VN, VC and* KN_3

The investigation of these structures (Peterson and Levy 1952b) was undertaken mainly for the purpose of determining the V and N coherent scattering cross sections. Both VN and VC have a rock-salt type of structure and, when absolute values of $F^2/16$ were plotted against $\sin^2\theta$, as in fig. 11, it

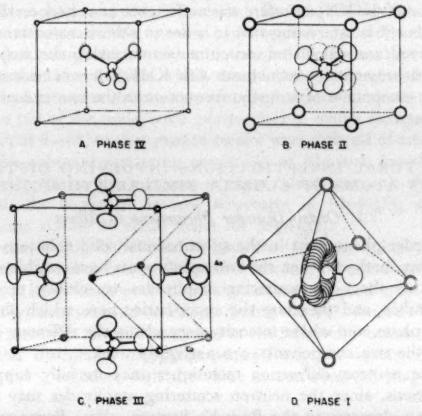


Fig. 10. Drawings of the structures of ND_3Br as determined by neutron diffraction. The circles and ellipses represent approximately the extent of thermal motion of the nuclei. The dotted lines in phase III show the pseudo-cubic cell. In phase II one ND_4^+ orientation is shown in full, the other with dotted lines. In phase I only one possible orientation is shown of one of the three possible models. (*After* Levy and Peterson 1953)

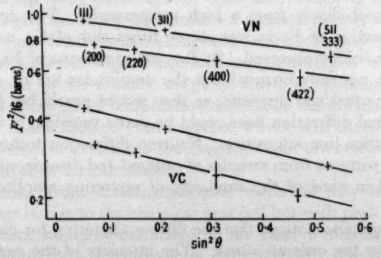


Fig. 11. A plot of observed neutron diffraction values of $F^2/16$ against $\sin^2\theta$ for vanadium nitride and vanadium carbide. The vertical lines indicate the estimated precision of the individual values. (*After* Peterson and Levy 1952b)

was clear that 'all odd' structure factors were systematically greater than 'all even' values, and that therefore V must have an amplitude opposite in sign to N and C, which are known to be positive. The absolute values of the V and N scattering cross sections are also obtained from the curves, and that of C is used as a check. Great care had to be taken to ensure the purity of the specimens used.

This negative value of the V scattering amplitude conflicts with the positive result obtained by McReynolds and Weiss (1951) in a total reflection experiment using a vanadium mirror. This may have been due to the presence of a surface oxide film on the mirror. Calculation (Hamermesh and Muehlhause 1950) using data on the 2700 ev scattering resonance in vanadium favours the negative sign.

Examination of the KN_3 structure was undertaken as a check on the N coherent cross section, but it is interesting that in order to achieve satisfactory agreement between observed and calculated structure factors, using the structure known from x-ray measurements (isostructural with KHF_2), it was necessary to assign an asymmetric component of temperature motion to the two end nitrogen atoms of the N_3^- group.

§7. STRUCTURAL INVESTIGATIONS INVOLVING DISTINCTION BETWEEN ATOMS OF CLOSELY SIMILAR ATOMIC NUMBER

7.1. *Order–Disorder Phenomena in Alloys*

Order–disorder phenomena in the substitutional solid solutions are difficult to study by x-ray methods when the two substituents have neighbouring atomic numbers, because the x-ray scattering amplitudes are closely proportional to the atomic number, and therefore the super-lattice lines which should appear in the ordered phase, and whose intensity depends on the *difference* in scattering amplitudes of the two components, are negligibly weak.

In this case neutron diffraction techniques may usefully supplement the x-ray investigations, since the neutron scattering amplitudes may vary widely for neighbouring elements in the Periodic System. Nix, Beyer and Dunning (1940) used this fact in an early study of the effect of heat treatment on the transmission of neutrons from a Ra–Be source through Fe–Ni samples. More recently Shull and Siegel (1949) have investigated differences in the neutron diffraction patterns of FeCo, Ni_3Mn and Cu_3Au alloys respectively, quenched and allowed to cool slowly from a high temperature. The neutron coherent scattering cross section of Fe is over eleven times that of Co, and super-lattice lines are therefore readily observed. So they are in the case of Ni_3Mn, where the cross sections are not only unequal but the amplitudes are of opposite phase. If they had been equal and opposite, as they would nearly be, for example, in $Mn^{60}Ni$, the normal diffraction lines could be made vanishingly weak compared with the super-lattice line intensities. Neutron diffraction techniques show no differences in the patterns from samples of ordered and disordered Cu_3Au, which is to be expected in view of the similarity of scattering amplitude of the two nuclei (table 1).

Shull and Siegel also noticed that the diffuse scattering for disordered alloys was larger than for the ordered alloys. The intensity of the super-lattice lines was used as a measure of the degree of ordering.

7.2. *Spinel Structures*

Another type of investigation requiring a differentiation between atoms is that of distinguishing between 'normal' and 'inverse' spinel structures, XY_2O_4 and $Y(X,Y)O_4$. In the 'normal' structure eight divalent metal ions are tetrahedrally co-ordinated and sixteen trivalent metal ions are octahedrally co-ordinated by O ions, which are themselves in the close-packed cubic positions. In the 'inverse' structure eight Y ions are tetrahedrally co-ordinated in the unit cell, while the remaining eight Y and the eight X ions are randomly arranged in sixteen octahedrally co-ordinated positions. To distinguish these structures by experiment the scattering powers of X and Y must be sufficiently different. This is not the case for the mineral spinel itself, $MgAl_2O_4$, if x-rays are used, but it is possible using neutron diffraction. Bacon (1952b) has shown that the structure is almost certainly 'normal', although it is possible that it is not quite ideal, but that a small degree of randomness of the inverse type is present. The intensity of the 111 reflection, which should show the largest difference between the normal and inverse structures, was only 5 neutrons/minute on a background of 50 at the tail of the undeviated neutron beam, which rises to a peak count of 500 000/min at $\theta = 0°$, so that precise results were difficult of attainment. A new value of the coherent scattering amplitude of Mg was found from cast Mg, MgO and Mg_2Sn for the purposes of these calculations. It should be possible to establish the supposed inverse structures of $MnFe_2O_4$ and $CoFe_2O_4$ by measurements similar to those made for $MgAl_2O_4$.

Magnetite, Fe_3O_4, is believed to have the 'inverse' spinel type of structure, the octahedral Fe sites being occupied at random by equal numbers of divalent and trivalent ions, whereas tetrahedral sites are occupied only by trivalent ions. Theoretically, x-rays should be able to distinguish Fe^{2+} from Fe^{3+}, but in practice this is not possible. It will be shown in the following section that it is possible to make this distinction using the neutron *magnetic* scattering amplitudes for Fe^{2+} and Fe^{3+}, together with the value of the observed magnetization.

§8. NEUTRON DIFFRACTION IN MAGNETIC MATERIALS

As already stated in a previous section, atoms having a resultant magnetic spin moment interact strongly with neutrons, and the observations of magnetic neutron scattering from magnetic materials may be classified under six headings:

(1) Scattering from paramagnetic materials, which is entirely incoherent, can be used to give the radial distribution of electron density in the 3d or 4f shell.

(2) Scattering from antiferromagnetic materials may show super-lattice peaks, the intensities of which vary with temperature and which disappear at a temperature near to the antiferromagnetic Curie temperature. The intensities of these peaks at any given temperature depend on the orientation of the opposed spins relative to the incident and scattered waves, and therefore, in principle at least, it is possible to determine the magnetic spin orientation of the ions either by trial and error or by calculation.

(3) Scattering from antiferromagnetic materials may indicate a magnetic unit cell similar in size to the chemical unit cell but with peak intensities different from those which would be calculated on the basis of nuclear scattering alone.

(4) Scattering from ferromagnetic materials will always show a magnetic unit cell similar to the chemical unit cell, but the intensity of the peaks, which is

the sum of the intensities of nuclear and magnetic scattering, will be dependent upon the strength and direction of any external magnetic field applied to the material.

(5) Transmission through magnetic materials depends on the scattering, coherent and incoherent, within them, and on their state of magnetization.

(6) Low-angle reflection from magnetic materials is again dependent on their magnetization.

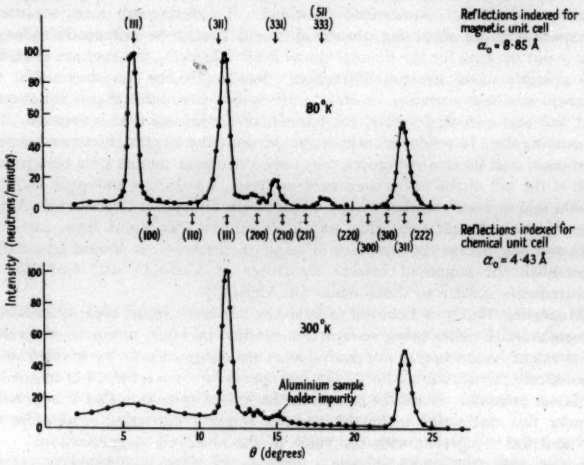


Fig. 12. Neutron diffraction patterns for MnO at liquid nitrogen and room temperatures, after correction for the various forms of extraneous diffuse scattering. Four extra antiferromagnetic reflections are to be noticed in the low-temperature pattern.
(*After* Shull, Strauser and Wollan 1951)

8.1. *Paramagnetic Materials* (Shull, Strauser and Wollan 1951)

These have been referred to mainly in an earlier section, but it is of interest that the diffuse scattering for neutrons of MnO at room temperatures shows a definite hump, similar to the scattering curve of x-rays by liquids. This implies that the magnetic moments are not quite randomly arranged, but that there is a partial short-range magnetic order.

8.2. *Antiferromagnetic Materials having Spin Moments only*

MnO is known to show a specific heat anomaly at 120°K, and is believed to be antiferromagnetic below that temperature. Neutron diffraction powder diagrams taken at 300°K and at 80°K are illustrated in fig. 12. They show the expected nuclear scattering pattern at 300°K typical of an NaCl-type of structure having nuclei scattering in opposite phase (the amplitude for Mn is negative), with a lattice constant of 4·426 Å. This pattern remains unchanged, though slightly enhanced in intensity, at liquid nitrogen temperatures, but there is superimposed on it a

super-lattice pattern corresponding to a cubic unit cell of twice the linear size, having $a_0 = 8{\cdot}85$ Å. For this cell only ' all odd ' reflections are observed, and the intensities fall off according to the same magnetic form factor deduced from the paramagnetic diffuse scattering.

Of the various magnetic structure models tested only that shown in fig. 13 gave satisfactory agreement with the observed data and with the results of precision x-ray measurements which indicate a slight distortion of the cubic framework (Tombs and Rooksby 1950). The magnetic lattice consists of sheets corresponding to a single set of (111) planes, within each of which the Mn ions are coupled ferromagnetically, but with antiferromagnetic coupling between neighbouring sheets. Along the cube axes there is always antiferromagnetic coupling from one Mn ion to the next, and this coupling persists through, or is caused by, an intermediate oxygen ion. This fits well with the phenomenon of *superexchange* advanced by Kramers (1934) and discussed in more detail by Anderson (1950). It should be noted that only two out of the possible eight {111} reflections contribute to the magnetic 111 reflection, whereas all eight contribute to the nuclear reflections.

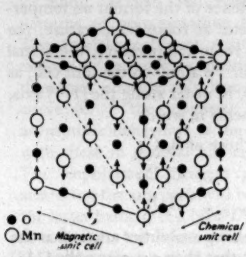


Fig. 13. The magnetic structure of MnO as shown by neutron diffraction at low temperatures. The arrows indicate the resultant spin moment orientations.
(By courtesy of *Science Progress*)

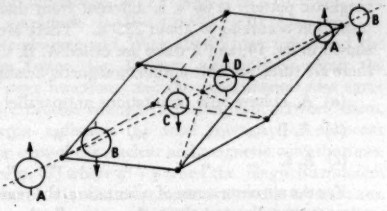


Fig. 14. Magnetic structure of α-Fe_2O_3.
(*After* Shull, Strauser and Wollan 1951)

MnS and MnSe give magnetic structures entirely similar to those of MnO. In all three the orientations of the opposed spin moments are calculated to be along the cube directions [100] rather than perpendicular to or parallel to the ferromagnetic (111) sheets.

The temperature variation of magnetic peak intensity, measured on the 111 super-lattice reflection for MnO, indicates that the structure perfection increases as the temperature is lowered. In the peak for 124°K, which is near to the antiferromagnetic Curie point, there is evidence of line broadening leading to a magnetic crystallite size of 50 Å.

MnF_2 has been studied extensively by Erickson (1952), who has shown that in this body-centred tetragonal structure (rutile-type) the magnetic unit cell is identical with the chemical unit cell, but the Mn^{2+} ion at $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$ has its spin antiparallel to that at 0, 0, 0, the directions of the spins being along the [001] direction. Intensity measurements were made down to 10·9° K.

8.3. *Antiferromagnetic Materials having Spin and Orbital Moments*

Shull, Strauser and Wollan (1951) have also studied the neutron diffraction patterns of FeO and CoO at liquid nitrogen and room temperatures and of NiO at room temperature, the Curie temperature in the last case being 250°C. The antiferromagnetic patterns for all these cubic oxides show the characteristic double-size magnetic cell with reflections observable only for 'all-odd' indices, though of course with different intensities from substance to substance. The weakness of the 111 magnetic reflection for FeO indicates that in this structure the magnetic moments of the Fe^{2+} ions are aligned perpendicular to the (111) ferromagnetic sheets, although NiO and CoO appear to be similar to MnO in moment alignment.

Shull, Strauser and Wollan have discussed in some detail the effect of the crystalline field in wholly or partially quenching the orbital moment and have calculated the magnetic cross sections for the various oxides for varying $(\sin\theta)/\lambda$ values, for comparison with experiment.

They have also studied α-Fe_2O_3 (haematite) over the range 80°K to 1000°K. The magnetic unit cell is of the same size as the chemical unit cell, which is rhombohedral with $a_0 = 5{\cdot}42$ Å, $\alpha = 55°17'$, $Z = 2$. It is possible to sort out the magnetic from the nuclear reflections by the dependence of the former on temperature. Although this substance is antiferromagnetic at room temperature, the magnetic pattern at 80°K is different from that at room temperature, the actual transition occurring at about 253°K. There are four Fe atoms in the unit cell, as shown in fig. 14, and if these are called A, B, C, D, in order along the [111] axis, there are three possible antiferromagnetic arrangements in which

(*a*) A, D have spin orientations antiparallel to those of B, C;

(*b*) A, B „ „ „ „ „ „ „ C, D;

(*c*) A, C „ „ „ „ „ „ „ B, D.

For the absolute sense of orientation, the reasonable possibilities are (i) that the moments are directed along the unit cell edges, (ii) that they are along the [111] direction, (iii) that they are in the (111) plane and directed towards one of the three nearest neighbours in that plane.

Structure factor calculations compared with experiment indicate that at room temperatures model (*a*) with orientation (iii) is correct, but that at low temperatures the structure is that of model (*a*) with orientation (ii). There are sheets of oxygen ions between each two antiferromagnetically coupled (111) sheets of Fe ions.

Erickson (1952) has found that FeF_2 and CoF_2 are essentially similar in magnetic structure to MnF_2, but that NiF_2 is different. In the latter the alignment of the magnetic moment appears to be at 10·0° with the tetragonal axis, probably along one of the two F–Ni–F directions.

Erickson has also studied polianite, MnO_2, in the metal ion of which the 3d orbital is less than half-filled. The magnetic unit cell has twice the linear dimensions of the chemical unit cell and the magnetic data, together with single crystal susceptibility measurements by Bizette and Tsai (1949), imply that the spin moments are along ⟨110⟩, normal to the tetragonal axis, and are antiparallel along all axial directions. Successive sheets parallel to (001) have their spin moments respectively along [110] (and $[\bar{1}\bar{1}0]$) and along $[1\bar{1}0]$ (and $[\bar{1}10]$). There is an indication that the magnetic contribution from the unquenched angular momentum is opposite to that from the electron spin, as might be expected for such an ion.

Erickson finds that for all the substances examined there appears to be a short-range order effect near the Curie temperature which diminishes quite slowly with temperature above T_c, indicating that the Curie temperature is not sharply defined, and that at the lowest temperature there is a residual diffuse magnetic scattering that is proportional to $\mathscr{S}$, the spin of the ion.

The observed neutron diffraction data for CoF_2 and NiF_2 give a sure indication of antiferromagnetism which susceptibility measurements do not in this case reveal. It is suggested that the latter should be repeated in as small a magnetic field as possible, since the effect of an external magnetic field on an antiferromagnetic substance, though formally zero, is not in fact fully understood.

Snow (1952) has investigated the magnetic structure of the antiferromagnetic compound CrSb, and found that the unit cell is unchanged. The chemical structure is similar to that of NiAs but with $a_0 = 4{\cdot}127$, $c_0 = 5{\cdot}451$ Å. The magnetic intensities are consistent with an arrangement of atomic magnetic moments along [001] so that alternate (001) planes are antiparallel, each plane being ferromagnetic in itself.

8.4. *Ferro- and Ferrimagnetic Materials*

Shull, Wollan and Koehler (1951) have studied the powder diffraction patterns of polycrystalline Fe and Co, using neutrons of wavelength 1·204 Å. The nuclear coherent scattering amplitude of Fe is $+0{\cdot}96 \times 10^{-12}$ cm; the expected magnetic scattering amplitude is $D = \pm 0{\cdot}598 f \times 10^{-12}$ cm and is, therefore, considerably less. Steinberger and Wick (1949) have calculated the curve for f, the magnetic form factor, for different assumptions about the asymptotic behaviour of the 3d wave functions, and the experimental data agree with these in general but are not precise enough to distinguish between them. For unpolarized incident neutron radiation the total intensity of coherent scattering for any peak will be the sum of the nuclear and magnetic contributions, the latter being given by q^2D^2 (see p. 17) where $q^2 = \frac{2}{3}$ when the magnetic moment of a domain has equal probability of all orientations relative to the crystal axes. In Fe the directions of easy magnetization of a single crystal are the cube edges, and the moments of separate domains also tend to be aligned along these directions, but the value of q^2 for all (hkl) planes is still $\frac{2}{3}$ for a polycrystalline, polydomain sample. It is not possible therefore to deduce the directions of the moments of individual domains relative to the axes by measurements of magnetic scattering intensities for different planes for such specimens.

Samples of isotopically enriched iron were also studied, although the samples of enriched ^{54}Fe and ^{57}Fe were very small. It was possible to show that the magnetic scattering from ^{56}Fe is the same as that from normal iron, as would be expected. Assuming this to be true also for the other two isotopes, the nuclear scattering amplitudes of ^{54}Fe and ^{57}Fe could be deduced.

Cobalt was studied in the form of metallic filings, and here the magnetic scattering is larger compared with the nuclear scattering. A curve for the magnetic form factor was obtained which is closely similar to that for iron.

Magnetite, Fe_3O_4, is believed to be ferrimagnetic, that is, the moments of the metal ions are arranged both ferromagnetically and antiferromagnetically, but there is a balance of ferromagnetism. Fe_3O_4 has a spinel structure; the formula may be written $Fe^{2+}Fe_2^{3+}O_4$. It is, however, believed to be an 'inverse' structure in which the sixteen octahedral sites are occupied randomly by divalent and trivalent ions whereas the eight tetrahedral sites are occupied only by trivalent

ions. The suggestion has been made (Verwey and Heilmann 1947, Néel 1948) that the magnetic moments of the octahedral ions are coupled antiferromagnetically with those of the tetrahedral ions, thus giving a ferromagnetic resultant of the correct quantitative value. Shull, Wollan and Koehler have been able to prove both these postulates to be correct (fig. 15). An alternative structure would have explained the observed magnetization by having the eight Fe^{2+} ions coupled ferromagnetically and occupying only the tetrahedral sites, with the octahedral sites occupied by paramagnetic trivalent ions. Intensity calculations for such a model would, however, give a 111 peak eight times weaker than observed. No changes in the magnetic structure were observed at low temperatures. It has been suggested (Verwey, Haayman and Romeijn 1947) that the pronounced increase in resistivity observed on lowering the temperature

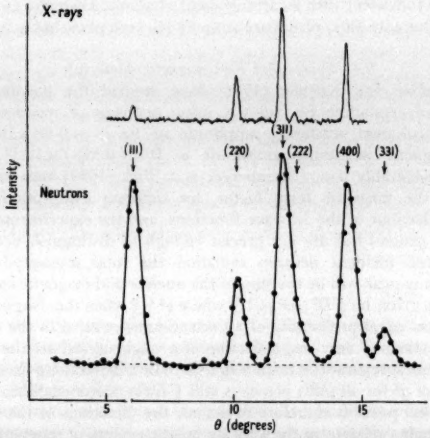


Fig. 15. Comparison between x-ray and neutron diffraction patterns for Fe_3O_4 (magnetite) at room temperature. Pronounced magnetic scattering contribution is found in the neutron pattern. (*After* Shull, Wollan and Koehler 1951)

through 120°K might be accounted for by an ordering of the Fe^{2+} and Fe^{3+} ions at the octahedral sites. The magnetic scattering amplitudes of Fe^{2+} and Fe^{3+} ($1{\cdot}088f\times10^{-12}$ and $1{\cdot}360f\times10^{-12}$) are not sufficiently different for such an ordering to be apparent from powder intensities where (*hkl*) peaks will tend to be superimposed, but it might be studied in single crystal, single domain specimens. This has been suggested but not yet carried out.

Tombs and Rooksby (1951) have pointed out that x-ray powder patterns show a change from cubic to rhombohedral symmetry in passing from 165°K to 95°K, and that this would be expected on the basis of antiparallel spins in successive (111) planes along one [111] direction only.

The variation of intensity with the application of an external magnetic field of various strengths up to saturation, in various directions relative to the scattering plane, has been dealt with in §2.3 (Shull, Wollan and Strauser 1951).

So also has the use of Fe_3O_4 for the production of a polarized neutron beam (Shull 1951), §2.4.

Bacon and Roberts (1953), using both the coherent magnetic intensities below the Curie point and the significant difference between the scattering amplitudes of Mg^{2+}, Al^{3+}, have obtained evidence in support of the incomplete inversion of $MgFe_2O_4$ and for the cationic distributions in magnesium ferrite aluminates which were suggested by Jones and Roberts (1952) from magnetic measurements. They also show that the relatively high scattering amplitude of oxygen leads to greater precision in the determination of the oxygen parameter u for ferrites, using neutron diffraction, than is possible with x-rays.

8.5. *Transmission of Neutrons through Magnetic Materials*

In an earlier section mention has been made of the method of producing a polarized neutron beam by transmission through highly magnetized samples of a ferromagnetic substance, usually pure iron, which possesses a different transmission cross section for the two characteristic spin states of the neutron. Hughes, Burgy and Woolf (1950) have measured the transmission of slow neutrons having velocities up to 4000 metres/second through a single crystal of iron with 10% Si content, when the crystal is placed in a position where Bragg scattering cannot occur. They have found that the incoherent scattering is largely inelastic, and the results support the calculation of polarization in polycrystalline iron by Steinberger and Wick (1949). Shull, Wollan and Koehler (1951) have measured the intensity variation around the 110 Debye ring obtained by using various blocks of iron of about one centimetre thickness as polarization analysers for a monochromatic incident neutron beam. This intensity variation is a measure of the grain orientation in the block. Burgy, Hughes, Wallace, Heller and Woolf (1950) have measured the change in polarization of a neutron beam with passage through unmagnetized iron and find that the rate of depolarization is a measure of the domain size. Such domain sizes have been measured for a number of samples of iron which have had different metallurgical treatments.

It is also interesting that the signs of the single transmission effect in the analysing block for the Fe (110) and the Fe_3O_4 (220) reflections are reversed, indicating that the directions of polarization in the two polarized beams are opposite. This is explained in terms of the magnetic structures of the two materials.

In the course of experiments on the production of polarized neutrons Hughes, Burgy, Heller and Wallace (1949) observed small-angle scattering (of the order of one minute) of neutrons in unmagnetized iron. Measurements of angular distribution and variation with iron thickness are consistent with the hypothesis that the scattering is caused by magnetic refraction of the neutrons at domain boundaries.

§9. OTHER APPLICATIONS

A number of investigations have been made which do not fall readily into the three main classes of application which we have discussed in the foregoing sections.

9.1. *Nuclear Positions*

We have already emphasized that neutrons determine the positions of nuclei rather than the clouds of electron density which are revealed by x-rays. Advantage has been taken of this in an investigation of the neutron diffraction

pattern of graphite by Bacon (1952a). It is shown that the positions of the carbon nuclei must be very close to those postulated by the Bernal (1924) structure and that the apparently anomalous x-ray intensities (Franklin 1950) are indeed due to limited knowledge of the electron distribution around the carbon nucleus. The present precision of neutron diffraction measurements is not, however, sufficient to distinguish any small departures from hexagonal symmetry in graphite such as have been proposed by Lukesh (1950) from single-crystal x-ray results.

9.2. *Diffraction by Liquids*

Chamberlain (1950) has studied the diffraction of neutrons in liquid sulphur, lead and bismuth using a divergent beam collimated by a single slit. In this experiment $(\sin\theta)/\lambda$ is well defined, but not θ or λ separately. The scattering sample was thin, but of large lateral dimensions, in a thin-wall aluminium cell. The results obtained were in agreement with those given by x-ray diffraction methods in showing that S has two nearest neighbours at about 2·1 Å, Pb has twelve at about 3·4 Å, and Bi an average of eight within 4 Å, but that there is no sharp differentiation between nearest and next-nearest neighbours.

9.3. *Diffraction by Gases*

A certain amount of information concerning coherent scattering amplitudes, and also about molecular structure, has been obtained from studies of diffraction by gases. In some early measurements with a beam of filtered neutrons, of mean wavelength 5·1 Å, Fermi and Marshall (1947) measured the total cross section of a number of gases such as CO_2, CF_4, N_2O and deduced, in particular, that the scattering amplitudes of all the component atoms in these compounds had the same sign.

Alcock and Hurst (1949, 1951) and Hurst and Alcock (1951) have made measurements of the angular distribution of neutrons scattered by D_2, O_2, CO_2, N_2, CF_4 and CH_4 using a monochromatic beam of wavelength 1·063 Å. The deuterium results were analysed by the complete quantum mechanical treatment of Hamermesh and Schwinger (1946), giving the information on the two scattering lengths a_2, a_4 for deuterium already referred to in §4. For the heavier molecules strict calculation covering all the possible transitions which may be undergone, under excitation by the neutron, becomes impracticable. However, Alcock and Hurst (1951) have shown that a semi-classical treatment, such as is normally used for x-ray scattering (Compton and Allison 1935, Pirenne 1946), in which the molecule is regarded as a rigid system of point scatterers, gives results of good accuracy in many cases. With nitrogen, for example, they found that S must be at least 90% of the value of σ. This at the time was inconsistent with deductions from crystal diffraction, but later diffraction measurements (Peterson and Levy 1952b) have supported the conclusions drawn from the gas scattering. From the observations with CF_4 it was concluded that σ, S for fluorine were almost equal and that the C–F bond length was 1·33 Å compared with 1·36 Å deduced by electron diffraction.

ACKNOWLEDGMENTS

The authors are indebted to Drs. C. G. Shull, S. W. Peterson, H. A. Levy and R. A. Erickson of Oak Ridge, to Mr. R. D. Lowde of Harwell and to the American Chemical Society for the use of material prior to publication.

LIST OF SYMBOLS

A mass number of nucleus
A_{hkl} absorption factor for hkl reflection
a scattering length of a nucleus with zero spin (bound unless specified)
a_+, a_- scattering lengths for parallel, antiparallel neutron spin orientations for a nucleus possessing spin
$\bar{a}$ coherent scattering amplitude of an element, or isotope, the average being taken over all relevant spin combinations and isotopes
a_0, b_0, c_0 dimensions of unit cell
α angle between the scattering and magnetization vectors
β monoclinic angle (obtuse)
B Debye factor in $\exp\{-2B(\sin^2\theta)/\lambda^2\}$
B_s magnetic induction in a domain
$\mathbf{c}$ velocity of light
$\Gamma_n^{(e)}$ width of resonance for re-emission of neutron with original energy
$\Gamma_a^{(e)}$ width of resonance for absorption
γ neutron magnetic moment in nuclear magnetons
δ difference of refractive index from unity, i.e. $n=1-\delta$
d interplanar spacing
D magnetic scattering amplitude (including form factor)
$-\mathbf{e}$ electronic charge
$\boldsymbol{\epsilon}$ unit scattering vector (in direction perpendicular to effective reflection planes)
E neutron energy
E_r resonance energy
$E_{el}(S)$ ordered elastic scattering cross section
$E_{el}(s)$ disordered ,, ,, ,, ,,
$E_{in}(S)$ ordered inelastic ,, ,, ,, ,,
$E_{in}(s)$ disordered ,, ,, ,, ,,
} functions varying with wavelength and temperature
f_x atomic scattering factor (amplitude) for X-rays
f magnetic form factor (amplitude) for neutrons
F_{hkl} structure factor (amplitude) for hkl reflection
ζ half-width of angular range of complete reflection for a perfect crystal
η standard deviation of angular spread of mosaic blocks
θ Bragg angle
θ_c critical value of θ for total internal reflection
h, k, l Miller indices of a reflection
$\boldsymbol{h}$ Planck's constant
$\mathcal{I}$ nuclear spin
I intensity of neutron beam (per minute per cm^2)
j_{hkl} number of co-operating planes for hkl reflection
$\boldsymbol{k}$ Boltzmann's constant
κ wave number
$\mathbf{K}$ unit vector in direction of atomic magnetic moment
λ wavelength
λ_c critical value of λ for total reflection
$\boldsymbol{\lambda}$ unit vector in neutron polarization direction
L counter slit length
$\boldsymbol{m}$ electron mass

m neutron mass
μ linear absorption coefficient
n refractive index
N number of nuclei per cm^3
N_c number of unit cells per cm^3
p_+, p_- relative weights of parallel, antiparallel spin orientations
p_r relative weight of any particular compound nucleus of scattering length a_r, allowing for spin and isotopes
p gas pressure
P neutron beam power (neutrons per minute)
$\mathbf{q}$ magnetic interaction vector
Q the crystallographic quantity $N_c^2\lambda^3F^2/\sin 2\theta$
ρ density
r distance
R nuclear radius
R_s radius of cylindrical specimen
$\mathscr{R}^\theta$ integrated reflection (rotating crystal method)
σ_t total cross section, i.e. absorption plus scattering
σ_a absorption cross section
σ_p polarization cross section
σ total scattering cross section
S coherent scattering cross section
s incoherent scattering cross section
} fundamental constants for isotopes and elements
σ_{eff} measured cross section in a transmission experiment
$d\sigma_m$ magnetic scattering cross section per unit solid angle
$\mathscr{S}$ spin quantum number of a magnetic atom
t thickness
τ total number of particles traversed
T absolute temperature
T_C Curie temperature
v velocity
V volume
ϕ phase difference between path through particle and the same distance in vacuum
χ angle between vibration direction of incident wave and direction of scattering
ω angular width of neutron beam
x, y, z coordinates of any point in the unit cell, expressed as fractions of a_0, b_0, c_0
Z number of molecules per unit cell

REFERENCES

Adair, R. K., 1950, *Rev. Mod. Phys.*, **22**, 249.
Alcock, N. Z., and Hurst, D. G., 1949, *Phys. Rev.*, **75**, 1609; 1951, *Ibid.*, **83**, 1100.
Anderson, P. W., 1950, *Phys. Rev.*, **79**, 350.
Arnold, G. P., Myers, V. W., and Weber, A. H., 1949, *Phys. Rev.*, **75**, 217.
Bacon, G. E., 1951, *Proc. Roy. Soc.* A, **209**, 397; 1952 a, *Acta Crystallogr.*, **5**, 492; 1952 b, *Ibid.*, **5**, 684.
Bacon, G. E., and Lowde, R. D., 1948, *Acta Crystallogr.*, **1**, 303.
Bacon, G. E., and Roberts, F. F., 1953, *Acta Crystallogr.*, **6**, 57.
Bacon, G. E., Smith, J. A. G., and Whitehead, C. D., 1950, *J. Sci. Instrum.*, **27**, 330.

BACON, G. E., and THEWLIS, J., 1949, *Proc. Roy. Soc.* A, **196**, 50.

BARNES, W. H., 1929, *Proc. Roy. Soc.* A, **125**, 670.

BENDT, P. J., and RUDERMAN, I. W., 1950, *Phys. Rev.*, **77**, 575.

BERNAL, J. D., 1924, *Proc. Roy. Soc.* A, **106**, 749.

BERNAL, J. D., and FOWLER, R. H., 1933, *J. Chem. Phys.*, **1**, 515.

BERNSTEIN, W., and SCHARDT, A. W., 1952, *Phys. Rev.*, **85**, 919.

BIZETTE, H., and TSAI, B., 1949, *Bull. Soc. Chim. Fr.*, [V] **16**, D 531. Langevin).

BLOCH, F., 1936, *Phys. Rev.*, **50**, 259; 1937, *Ibid.*, **51**, 994.

BRADLEY, A. J., 1935, *Proc. Phys. Soc.*, **47**, 879.

BREIT, G., and WIGNER, E., 1936, *Phys. Rev.*, **49**, 519.

BRILL, T., and LICHTENBERGER, H. V., 1947, *Phys. Rev.*, **72**, 585.

BURGY, M., HUGHES, D. J., WALLACE, J. R., HELLER, R. B., and WOOLF, W. E., 1950, *Phys. Rev.*, **80**, 953.

BURGY, M. T., RINGO, G. R., and HUGHES, D. J., 1951, *Phys. Rev.*, **84**, 1160.

CASSELS, J. M., 1950, *Progress in Nuclear Physics*, I, ed. O. R. Frisch (London : Butterworth-Springer); 1951, *Proc. Roy. Soc.* A, **208**, 527.

CHAMBERLAIN, O., 1950, *Phys. Rev.*, **77**, 305.

CLAASSEN, A., 1930, *Phil. Mag.*, **9**, 57.

COMPTON, A. H., and ALLISON, S. K., 1935, *X-rays in Theory and Experiment* (New York : Van Nostrand).

DARWIN, C. G., 1914, *Phil. Mag.*, **27**, 315, 675.

DEBYE, P., 1913, *Verh. dtsch. phys. Ges.*, **15**, 678, 738, 857; 1914, *Ann. Phys., Lpz.*, **43**, 49.

EGELSTAFF, P., 1951, *Nature, Lond.*, **168**, 290.

EKSTEIN, H., 1949, *Phys. Rev.*, **76**, 1328; 1950, *Ibid.*, **78**, 731.

ELSASSER, W. M., 1936, *C. R. Acad. Sci., Paris*, **202**, 1029.

ERICKSON, R. A., 1952, *Thesis*, University of Texas.

FANKUCHEN, I., 1937, *Nature, Lond.*, **139**, 193.

FERMI, E., 1950, *Nuclear Physics*, notes compiled by J. Orear, A. H. Rosenfeld and R. A. Schluter. Revised edn. (Chicago : University of Chicago Press).

FERMI, E., and MARSHALL, L., 1947, *Phys. Rev.*, **71**, 666; 1949, *Ibid.*, **75**, 578.

FERMI, E., STURM, W. J., and SACHS, R. G., 1947, *Phys. Rev.*, **71**, 589.

FERMI, E., and ZINN, W. H., 1946, *Phys. Rev.*, **70**, 103.

FESHBACH, H., PEASLEE, D. C., and WEISSKOPF, V. F., 1947, *Phys. Rev.*, **71**, 145.

FINKELSTEIN, R. J., 1947, *Phys. Rev.*, **72**, 907.

FLEEMAN, J., NICODEMUS, D. B., and STAUB, H. H., 1949, *Phys. Rev.*, **76**, 1774.

FOWLER, I. L., and TUNNICLIFFE, P. R., 1950, *Rev. Sci. Instrum.*, **21**, 734.

FRANKLIN, R, E., 1950, *Nature, Lond.*, **165**, 71.

FRIED, B. D., 1952, *Phys. Rev.*, **86**, 434.

GANS, R., 1925, *Ann. Phys., Lpz.*, **76**, 29.

GOLDSCHMIDT, G. H., and HURST, D. G., 1951, *Phys. Rev.*, **83**, 88; 1952, *Ibid.*, **86**, 797.

GOLDSMITH, H. H., IBSER, H. W., and FELD, B. T., 1947, *Rev. Mod. Phys.*, **19**, 259.

GORDON, M. M., 1950, *Phys. Rev.*, **80**, 1111.

GUTOWSKY, H. S., and PAKE, G. E., 1948, *J. Chem. Phys.*, **16**, 1164.

HALBAN, H., and PREISWERK, P., 1936, *C. R. Acad. Sci., Paris*, **203**, 73.

HALPERN, O., 1949, *Phys. Rev.*, **76**, 1130.

HALPERN, O., HAMERMESH, M., and JOHNSON, M. H., 1941, *Phys. Rev.*, **59**, 981.

HALPERN, O., and HOLSTEIN, T., 1941, *Phys. Rev.*, **59**, 960.

HALPERN, O., and JOHNSON, M. H., 1939, *Phys. Rev.*, **55**, 898.

HAMERMESH, M., 1942, *Phys. Rev.*, **61**, 17.

HAMERMESH, M., and MUEHLHAUSE, C. O., 1950, *Phys. Rev.*, **78**, 175.

HAMERMESH, M., RINGO, G. R., and WATTENBERG, A., 1952, *Phys. Rev.*, **85**, 483.

HAMERMESH, M., and SCHWINGER, J., 1946, *Phys. Rev.*, **69**, 145.

HAVENS, W. W. Jr., RAINWATER, L. J., WU, C. S., and DUNNING, J. R., 1948, *Phys. Rev.*, **73**, 963.

HIBDON, C. T., and MUEHLHAUSE, C. O., 1950, *Phys. Rev.*, **79**, 44.

HIBDON, C. T., MUEHLHAUSE, C. O., SELOVE, W., and WOOLF, W., 1950, *Phys. Rev.* **77**, 730.

HOOVER, J. I., JORDAN, W. H., MOAK, C. D., PARDUE, L., POMERANCE, H., STRONG, J. D., and WOLLAN, E. O., 1948, *Phys. Rev.*, **74**, 864.
HUGHES, D. J., and BURGY, M. T., 1951, *Phys. Rev.*, **81**, 498.
HUGHES, D. J., BURGY, M. T., HELLER, R. B., and WALLACE, J. W., 1949, *Phys. Rev.*, **75**, 565.
HUGHES, D. J., BURGY, M. T., and RINGO, G. R., 1950, *Phys. Rev.*, **77**, 291.
HUGHES, D. J., BURGY, M. T., and WOOLF, W. E., 1950, *Phys. Rev.*, **80**, 481.
HUGHES, D. J., WALLACE, J. R., and HOLTZMAN, R. H., 1948, *Phys. Rev.*, **73**, 1277.
VAN DE HULST, H. C., 1946, *Optics of Spherical Particles* (Amsterdam : Duwaer).
HUNT, E. B., and RUNDLE, R. E., 1951, *J. Amer. Chem. Soc.*, **73**, 4777.
HURST, D. G., and ALCOCK, N. Z., 1951, *Canad. J. Phys.*, **29**, 36.
HURST, D. G., PRESSESKY, A. J., and TUNNICLIFFE, P. R., 1950, *Rev. Sci. Instrum.*, **21**, 705.
International Tables for the Determination of Crystal Structures, 1935 (Berlin : Borntraeger).
JAMES, R. W., 1948, *The Opticol Principles of the Diffraction of X-rays* (London : Bell).
JONES, G. O., and ROBERTS, F. F., 1952, *Proc. Phys. Soc.* B, **65**, 390.
KLEINMAN, D., and SNOW, G., 1951, *Phys. Rev.*, **82**, 952.
KOEHLER, W. C., and WOLLAN, E. O., 1952, *Phys. Rev.*, **85**, 491.
KRAMERS, H. A., 1934, *Physica*, **1**, 182.
LAX, M., 1950, *Phys. Rev.*, **80**, 299.
LEVY, H. A., and PETERSON, S. W., 1951, *Phys. Rev.*, **83**, 1270; 1952, *Phys. Rev.*, **86**, 766; 1953, *J. Amer. Chem. Soc.*, **75**, 1536.
LONSDALE, K., 1943, *Rep. Prog. Phys.*, **9**, 256 (London : Physical Society); 1947, *Miner. Mag.*, **28**, 14.
LOWDE, R. D., 1950, *Rev. Sci. Instrum.*, **21**, 835; 1951, *Nature, Lond.*, **167**, 244; 1952, *Proc. Phys. Soc.* A, **65**, 857.
LUKESH, J. S., 1950, *Phys. Rev.*, **80**, 226.
MCDANIEL, B. D., SUTTON, R. B., LAVATELLI, L. S., and ANDERSON, E. E., 1947, *Phys. Rev.*, **72**, 729.
MCREYNOLDS, A. W., 1951, *Phys. Rev.*, **84**, 969.
MCREYNOLDS, A. W., and WEISS, R. J., 1951, *Phys. Rev.*, **83**, 171.
MITCHELL, D. P., and POWERS, P. N., 1936, *Phys. Rev.*, **50**, 486.
MOORHOUSE, R. G., 1951, *Proc. Phys. Soc.* A, **64**, 1097.
MOTT, N. F., and MASSEY, H. S. W., 1949, *The Theory of Atomic Collisions*, 2nd edn. (Oxford : University Press).
VON NARDROFF, R., 1926, *Phys. Rev.*, **28**, 240.
NÉEL, L., 1948, *Ann. Phys., Paris*, **3**, 137.
NIX, F. C., BEYER, H. C., and DUNNING, J. R., 1940, *Phys. Rev.*, **58**, 1031.
Nuclear Data, 1950 (Washington : National Bureau of Standards).
PAULING, L., 1935, *J. Amer. Chem. Soc.*, **57**, 2680; cf. also 1940, *Nature of the Chemical Bond*, 2nd edn. (Ithaca : Cornell University Press), p. 301.
PAULING, L., and EWING, F. J., 1948, *J. Amer. Chem. Soc.*, **70**, 1660.
PETERSON, S. W., and LEVY, H. A., 1951, *J. Chem. Phys.*, **19**, 1416; 1952 a, *Ibid.*, **20**, 704; 1952 b, *Phys. Rev.*, **87**, 462.
PIRENNE, M. H., 1946, *The Diffraction of X-rays and Electrons by Free Molecules* (Cambridge : University Press).
POMERANCE, H., 1951, *Phys. Rev.*, **83**, 641; 1952, *Ibid.*, **88**, 412.
RAINWATER, L. J., HAVENS, W. W. Jr., DUNNING, J. R., and WU, C. S., 1948, *Phys. Rev.*, **73**, 733.
RAYLEIGH, Lord, 1911, *Proc. Roy. Soc.* A, **84**, 25.
RUDERMAN, I. W., 1949, *Phys. Rev.*, **76**, 1572; 1950, *Ibid.*, **78**, 317(A).
RUNDLE, R. E., 1951, *J. Amer. Chem. Soc.*, **73**, 4172.
RUNDLE, R. E., SHULL, C. G., and WOLLAN, E. O., 1952, *Acta Crystallogr.*, **5**, 22.
SAWYER, R. B., WOLLAN, E. O., BERNSTEIN, S., and PETERSON, K. C., 1947, *Phys. Rev.*, **72**, 109.
SCHENCK, J., and HEATH, R. L., 1952, *Phys. Rev.*, **85**, 923.
SCHWINGER, J. S., 1937, *Phys. Rev.*, **51**, 544.
SHULL, C. G., 1951, *Phys. Rev.*, **81**, 626.
SHULL, C. G., and SIEGEL, S., 1949, *Phys. Rev.*, **75**, 1008.
SHULL, C. G., STRAUSER, W. A., and WOLLAN, E. O., 1951, *Phys. Rev.*, **83**, 333.

SHULL, C. G., and WOLLAN, E. O., 1948, *Science*, **108**, 69; 1951, *Phys. Rev.*, **81**, 527.
SHULL, C. G., WOLLAN, E. O., and KOEHLER, W. C., 1951, *Phys. Rev.*, **84**, 912.
SHULL, C. G., WOLLAN, E. O., MORTON, G. A., and DAVIDSON, W. L., 1948, *Phys. Rev.*, **73**, 842.
SHULL, C. G., WOLLAN, E. O., and STRAUSER, W. A., 1951, *Phys. Rev.*, **81**, 483.
SIDHU, S. S., 1948, *J. Appl. Phys.*, **19**, 639.
SNOW, A. I., 1952, *Phys. Rev.*, **85**, 365.
SQUIRES, G. L., 1952, *Proc. Roy. Soc.* A, **212**, 192.
STEINBERGER, J., and WICK, G. C., 1949, *Phys. Rev.*, **76**, 994.
STURM, W. J., 1947, *Phys. Rev.*, **71**, 757.
SUTTON, R. B., *et al.*, 1947, *Phys. Rev.*, **72**, 1147.
TOMBS, N. C., and ROOKSBY, H. P., 1950, *Nature, Lond.*, **165**, 442; 1951, *Acta Crystallogr.*, **4**, 474.
TROESCH, A., and VERDE, M., 1951, *Helv. Phys. Acta*, **24**, 39.
VERWEY, E. J., HAAYMAN, P. W., and ROMEIJN, F. C., 1947, *J. Chem. Phys.*, **15**, 181.
VERWEY, E. J. W., and HEILMANN, E. L., 1947, *J. Chem. Phys.*, **15**, 174.
WALLER, I., 1923, *Z. Phys.*, **17**, 398; 1925, *Dissertation*, Uppsala.
WEIGLE, J., and SAINI, H., 1936, *Helv. Phys. Acta*, **9**, 515.
WEINBERG, A. M., and SCHWEINLER, H. C., 1948, *Phys. Rev.*, **74**, 851.
WEINSTOCK, R., 1944, *Phys. Rev.*, **65**, 1.
WEISS, R. J., 1951, *Phys. Rev.*, **83**, 379; 1952, *Ibid.*, **86**, 271.
WEISS, R. J., HASTINGS, J., and CORLISS, L., 1951, *Phys. Rev.*, **83**, 863.
WHITAKER, M. D., 1937, *Phys. Rev.*, **52**, 384.
WHITAKER, M. D., BEYER, H. G., and DUNNING, J. R., 1938, *Phys. Rev.*, **54**, 771.
WHITAKER, M. D., and BRIGHT, W. C., 1940, *Phys. Rev.*, **57**, 1076 A; 1941, *Ibid.*, **60**, 280.
WINSBERG, L., MENEGHETTI, D., and SIDHU, S. S., 1949, *Phys. Rev.*, **75**, 975.
WOLLAN, E. O., DAVIDSON, W. L., and SHULL, C. G., 1949, *Phys. Rev.*, **75**, 1348.
WOLLAN, E. O., and SHULL, C. G., 1948 a, *Phys. Rev.*, **73**, 830; 1948 b, *Nucleonics*, **3**, 2, 17.
WOLLAN, E. O., SHULL, C. G., and KOEHLER, W. C., 1951, *Phys. Rev.*, **83**, 700.
WOLLAN, E. O., SHULL, C. G., and MARNEY, M. C., 1948, *Phys. Rev.*, **73**, 527.
ZACHARIASEN, W. H., 1945, *Theory of X-ray Diffraction in Crystals* (New York: Wiley).

PHYSICAL PROPERTIES AND ATOMIC ARRANGEMENTS IN CRYSTALS

By W. A. WOOSTER
Department of Mineralogy and Petrology, University of Cambridge

CONTENTS

PAGE

§1. Introduction 62
§2. Magnetic properties and atomic arrangements in crystals 63
2.1. Diamagnetic crystals 63
2.2. Paramagnetic crystals 64
2.3. Ferromagnetic crystals 65
§3. Optical properties and atomic arrangements in crystals 67
3.1. Optical activity 67
3.2. Raman spectra and infra-red absorption spectra 70
4. Piezoelectric properties and atomic arrangements in crystals 71
4.1. Normal piezoelectric crystals 72
4.2. Seignette-electric group 74
§5. Elastic properties and atomic arrangements in crystals 76
5.1. The relation of c_{11} to the nature of the atomic binding ... 77
5.2. Theories 80
References 81

ABSTRACT. A survey is given of the relation between four physical properties of crystals and their structural characters. The properties discussed are the magnetic, optical, piezoelectric and elastic. The well-established relations between (*a*) diamagnetic anisotropy and molecular orientation and (*b*) paramagnetic anisotropy and the environment of complex ions are briefly reviewed. The more recent study of paramagnetic resonance absorption at high frequencies in a magnetic field is also included. The work on antiferromagnetic crystals is discussed in connection with the spinels. The study of optical activity and Raman spectra are considered under the heading of optical properties. The experimental confirmation by x-rays of the original convention concerning atomic configurations and the sign of optical activity in a molecule are described and also the connection between special features of the structures of tartrates, of sodium chlorate and of quartz and the handedness of the rotatory power is analysed. The piezoelectric properties of zinc blende, quartz, tourmaline, Rochelle salt and potassium dihydrogen phosphate are considered in relation to their structures. In the study of quartz and tourmaline the origin of the piezoelectric charge is attributed to tetrahedra of SiO_4 stressed in a direction from one corner to the opposite face. The spontaneous electric polarization of Rochelle salt and KH_2PO_4 is attributed to cooperative movements of hydrogen ions. Broad generalizations concerning the relation between elastic constants and the crystal structure are given. It is shown that the hardness increases in a regular way with the elastic constant c_{11} (or an average of c_{11}, c_{22}, c_{33}) and also that there is a close connection between directions of strong interatomic bonds and a large value of the corresponding elastic constant.

§1. INTRODUCTION

THE ever-increasing knowledge about the atomic arrangements in crystals makes possible a better understanding of the origin of their distinctive physical properties. The attention devoted to the various physical properties varies notably from one decade to another. There has also been an uneven development of the various branches of crystal physics. This may be attributed to two causes. Technical applications of crystalline properties have

led to intensive experimental research in certain fields and modern theories relating to atomic structure and interatomic bonds have been found specially suitable for the study of a few of the physical properties.

In attempting to make a correlation between crystal structure and physical properties a fundamental difficulty arises. This is due to the fact that nearly all properties depend on several factors, such as the electronic energy levels of the atoms, the presence of valency bonds or Brillouin zones, which are only in part determined by the crystal structure. It is therefore not to be expected that there should in every property be an equally close correlation between structure and physical property. However, there is frequently some connection between the two and it is of great importance in leading to an understanding of the way in which any substance, having a desired physical property, can be prepared. There has been an enormous number of purely theoretical studies dealing with very simple atomic arrangements. Almost every property of the alkali halides, for example, has been studied on the basis of every theory, classical and modern. This concentration on a particular class of crystal may be justified by the difficulty of dealing with more complicated arrangements of atoms. However, it is not of much immediate help in understanding the physical properties of the vast crystal kingdom in which the majority of crystals are monoclinic. The approach to the subject which is employed here is based on the conception that if the physical properties of a large number of crystals of all types of symmetry are studied, certain general empirical relations are likely to emerge. Such rules form valuable guides in the further study of the actual crystals and may suggest fruitful lines for theoretical developments. The four properties to be discussed here are magnetic, optical, piezoelectric and elastic. This may seem a small number out of the dozen or so physical properties of crystals but they have been chosen because the work done in connection with them is largely recent and also because they provide good examples of the correlation under discussion.

§ 2. MAGNETIC PROPERTIES AND ATOMIC ARRANGEMENTS IN CRYSTALS

It is convenient to divide the discussion into sections dealing with (i) diamagnetic, (ii) paramagnetic and (iii) ferromagnetic crystals.

2.1. *Diamagnetic Crystals*

Since the early work of Faraday and Tyndall the study of diamagnetic crystals has played an important part in the development of techniques of measurement and the theories of magnetism. The great advance began after the discovery of the exact atomic arrangements in crystals because it then became possible to correlate the magnetic measurements with the geometric positions of the atoms (except where a particular reference is given the crystal structure is described in the *Strukturbericht*, Vol. I–VII, 1931 (Leipzig: Akad. Verlag)). There is a closer relation between this physical property and the crystal structure than is found with any other physical property. This is due to the very weak interaction between a diamagnetic material and the magnetic field. The magnetic dipoles induced by the field are of such a small magnitude that, with very few exceptions, each dipole is unaffected by the presence of the others. This leads at once to the possibility that the total magnetic effect of the crystal is the sum of the contributions of the atoms or molecules taken separately. Many organic

molecular crystals have been studied and it is found that three principal susceptibilities can be assigned to each molecule (Krishnan, Guha and Banerjee 1933, Krishnan and Banerjee 1935 a, b, Pauling 1936, Lonsdale and Krishnan 1936, Lonsdale 1938, 1939, Banerjee 1938). Employing the usual tensor treatment the total effect of the actual arrangement of the molecules is obtained by simple summation over a unit cell. This simple relationship made it possible to check, in crystals of suitable symmetry, the atomic arrangement determined by x-ray methods. If the molecule is of known form its principal susceptibilities can be calculated from atomic susceptibilities with reasonable accuracy (Pascal 1913). These values can then be used in calculating the principal susceptibilities of the crystal from the proposed relative positions of the molecules. The accuracy of this check is such that the magnetic and x-ray determinations of the orientation of the normal to a flat aromatic molecule do not usually differ by more than one or two degrees. In calculating the principal molecular susceptibilities it is necessary to take account of the special influence of rings, and various types of bonds (Pauling 1936, Lonsdale 1939). The additive relations for diamagnetic susceptibility put forward by Pascal (1913) have to be modified according to the double, triple or other special bonds present and above all according to the number and orientation of benzene rings present. Subject to these qualifications it may be said that the empirical relation between diamagnetic properties and atomic arrangement is fully investigated. There remains to be fully worked out the theoretical interpretation in terms of electronic and atomic theory of the actual values obtained by experiment.

2.2. *Paramagnetic Crystals*

Whereas diamagnetism is interpreted by the action of the applied magnetic field on the electron orbits of paired electrons and completed shells, paramagnetism is attributed to the action on the unpaired electrons occurring either in inner shells, as in the rare-earth elements, or in the outer shells, as in the transition elements. The arrangement of the atoms forming the immediate neighbourhood of a paramagnetic ion, in ionic crystals, determines the so-called crystalline field. This field may cause variations of the susceptibility with direction. One of the best studied examples of this is afforded by $CuSO_4.5H_2O$. Each Cu^{2+} ion is surrounded by four water molecules lying in a plane containing the Cu ion and two oxygen atoms lying further from the Cu ion than the water molecules on a line perpendicular to the plane containing the water molecules. In each unit cell there are two Cu ions and the planes of the water molecules are inclined almost at 90° to one another. The line of intersection of these planes is an axis of magnetic symmetry of the crystal and all directions perpendicular to this axis have the same susceptibility. Thus the magnetic property of the crystal is dominated by the contribution of the copper ion and this in turn is made anisotropic by the surrounding water molecules and oxygen ions (Krishnan, Chakravorty and Banerjee 1933, Krishnan and Mookherji 1936, 1937 c, 1938 a, b). A tetrahedral environment as, for instance, occurs in $Cs_2[CoCl_4]$, leads to a very small anisotropy of the paramagnetic susceptibility of the crystal (Krishnan and Mookherji 1937 a, b). In this substance the maximum variation from the mean susceptibility is only about 3%. In $CuSO_4.5H_2O$ the difference between the principal susceptibilities of the copper group are about 40% of the mean susceptibility. Tutton salts such as $(NH_4)_2SO_4 . FeSO_4 . 6H_2O$ show an anisotropy less than 1% and this corresponds to the octahedral environment of water molecules

surrounding the Fe ion. These examples serve to show how the nature of the environment of certain ions of transition metals is indicated by the paramagnetic anisotropy.

In recent years a measurement of another kind has been developed in connection with paramagnetic ions (Gorter 1936 a, b, Zavoisky 1945, Cummerow and Halliday 1946, Wheatley and Halliday 1949). These measurements determine the magnetic power loss in a crystal subjected simultaneously to an alternating magnetic field and a steady field perpendicular to the alternating field. The resonance frequency of the oscillation at which the important absorption of a magnetic energy occurs is the same as that of the Larmor precession of the ion, in the steady field. Thus at frequencies of the order of 10^{10} c/s the necessary steady field is of the order of 10^3 oersted and at lower frequencies the steady field is proportionally lower. The theory of this effect has been studied by a number of authors (Waller 1932, Frenkel 1945).

The relation between the frequency ν of the electromagnetic radiation and the strength of the applied magnetic induction B is given by the equation $h\nu = g\beta B$ where g is the gyromagnetic ratio, i.e. the ratio of the angular momentum to the magnetic moment of the ion, and β is the magnetic moment of the Bohr magneton. From this equation g can be determined. The properties of $CuSO_4 . 5H_2O$ have been studied by this method (Cummerow and Halliday 1946, Bagguley and Griffiths 1948, Wheatley and Halliday 1949) and at a micro-wavelength of 0·85 cm an exact correspondence within the experimental error between the g-value observed here and that calculated from the static measurements (Krishnan and Mookherji 1936, 1938 a, b) of susceptibility is obtained. At longer wavelengths of the electromagnetic radiation certain absorption bands are not quite resolved. The calculation from the static measurements involves the assumption that the orbital moment of the ion is quenched and that there is only a spin contribution to the total moment. This accords very well with the usual theory of the transition-metal paramagnetic-ion type.

Observations of the susceptibility of paramagnetic salts have lead to information concerning the Stark splitting of the various electronic energy levels but this work has so far not been correlated in a detailed manner with the atomic arrangements (Krishnan and Banerjee 1936, Krishnan and Mookherji 1938 c, Krishnan, Mookherji and Bose 1939).

2.3. *Ferromagnetic Crystals*

It is probably true to say that in recent years most progress in tracing the connection between atomic arrangement and ferromagnetic properties has been made in connection with those crystals which show antiferromagnetism. The first suggestion that there might be such a property seems to have been made to Serres (1932) by Néel. The theory that ferromagnetism is due to the mutual influence of atomic magnetic moments which tend to be oriented parallel to one another in a magnetic field has been accepted for a long time. Néel suggested that there might be antiferromagnetic materials in which the magnetic moments of neighbouring atoms are oppositely directed. The simplest example of such an arrangement would be one in which the atoms at the corners of a cubic unit cell have magnetic moments directed along [001] with the N pole uppermost while the atom at the centre of each cell has its magnetic axis also parallel to [001] but with the S pole uppermost (fig. 1). This theory affords an explanation of many

of the magnetic properties of the series of mixed oxides known as the ferrites (Néel 1948, 1950, Néel and Brochet 1950), of composition Fe_2O_3 . MO, [M = Mn, Co, Fe, Cu, Zn, Cd, Li]. These have the spinel structure in which certain metal atoms are in positions of fourfold coordination and others in sixfold coordination. The various ferrites can be classified into two groups called normal and inverse. In the normal group there are only Fe-atoms in sixfold positions and M-atoms in fourfold positions. The only examples of normal ferrites are ZnO . Fe_2O_3, and CdO . Fe_2O_3 which are paramagnetic. The inverse group has all the fourfold and half the sixfold positions filled with Fe atoms and the remainder of the sixfold positions filled with M atoms. M may be Mn, Fe, Co, Ni, Cu, Li and the crystals are in general ferromagnetic. Néel supposes that the normal group is not ferromagnetic because the mutual influence of the magnetic moments of the neighbouring atoms cannot come into operation. In this group each Fe-ion has four iron atoms near to it and the co-ordination polyhedron is nearly flat. In the inverse ferromagnetic group of ferrites it is postulated that only the M-type atoms contribute to the ferromagnetism, while the iron atoms in the four- and sixfold positions form an antiferromagnetic arrangement due to their relative positions in the structure. The observed susceptibilities of the ferrites are explained on this basis if the usual magneton numbers are associated with the M-type atoms.

The theory of antiferromagnetism has received support from two other types of observation. Using x-rays it has been shown that certain oxides which at room temperature are cubic pass to a lower symmetry at a transition point usually between 95° K and 300° K (Tombs and Rooksby 1950, 1951, Rooksby and Tombs 1951, Smart and Greenwald 1951). Thus CoO changes from face-centred cubic to tetragonal; FeO, NiO and MnO change to a rhombohedral unit cell. Magnetite Fe_2O_3 . FeO becomes rhombohedral between 165° K and 95° K. It is suggested that in the oxides which become rhombohedral the arrangement of the magnetic dipoles is as shown in fig. 2. At every lattice point of the cubic face-centred unit

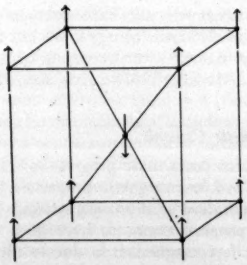


Fig. 1. Diagram showing an antiferromagnetic arrangement for a body-centred cubic lattice.

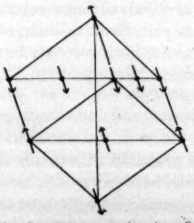


Fig. 2. Diagram showing how an antiferromagnetic arrangement for a face-centred cubic unit cell leads to a rhombohedral cell.

cell we suppose an Fe or Ni or Mn atom with its magnetic axis directed along one of the cube edges. Below the transition temperature consecutive (111) planes have atoms with oppositely directed moments and their interaction determines the rhombohedral character of the unit cell and the antiferromagnetic character of the crystal. The possibility of developing a tetragonal lattice due

to such magnetic interactions is shown by fig. 3. The departure of magnetite from cubic magnetic symmetry below the transition temperature of 130° K has been established by Bickford (1950) using the method of resonance absorption of micro-electromagnetic waves.

The other type of observation bearing on this transition has been made by diffracting neutrons from powdered MnO (Shull and Smart 1949). At 300° K the usual diffraction pattern is obtained but at 80° K the pattern is based on a new unit cell, still face-centred cubic, but having a side twice as long as that of the

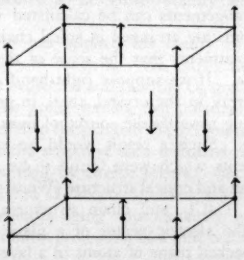


Fig. 3. Diagram showing how an antiferromagnetic arrangement for a face-centred cubic unit cell leads to a tetragonal cell.

higher-temperature form. The total scattering is provided about equally by the nuclear and magnetic spin scattering. In the antiferromagnetic state the magnetic spins are reversed as shown in fig. 2 and hence the 'magnetic' unit cell which is recorded by the neutrons is twice as large as the crystallographic cell. Owing to the relatively low resolution of the neutron 'lines' it is not possible to establish the rhombohedral nature of the structure.

§3. OPTICAL PROPERTIES AND ATOMIC ARRANGEMENTS IN CRYSTALS

In recent years there has been an enormous development of the knowledge concerning the optical properties of crystals. Studies have been made in the ultra-violet and infra-red regions of the spectrum; the change of optical properties with stress and strain have been investigated especially in connection with piezo-electric vibrations; the optical properties of metals and semi-metals have been investigated partly on account of the need to identify the many ore-minerals which cannot be studied in the ordinary optical microscope; and, lastly, there has been a great application of theories to explain in terms of fundamental quantities the manifold optical properties of crystals. Although nearly all of this work has some bearing on the relation between structure and optical properties only in two sections has there been an extensive addition to our knowledge of this relation. These sections are respectively concerned with optical activity and light scattering.

3.1. *Optical Activity*

Ever since the time of Fresnel (1825) attempts have been made to work out the relation between the atomic arrangement and the rotatory power in optically active materials. Born (1933) and Kirkwood (1937) have reviewed the subject.

Recent presentations of theories are due to Lévy (1950), Volkenshtein and Byutner (1951), Ramachandran (1951 a, b, c) and to Wood, Fickett and Kirkwood (1952). Some of these papers deal with optically active molecules isolated from one another while others deal with the arrangements of atoms found in crystals. A sharp distinction between the two approaches cannot be made because molecules preserve much of their individual characters when built into crystals. Broadly speaking the result of the numerous theories which have been put forward is to show that, accepting certain assumptions about the polarizability of the atoms composing the molecule or crystal, the strength of optical activity in not too complex atomic arrangements can be calculated with reasonable accuracy. In some crystals the atoms are arranged in spiral chains and it has been shown by several theoretical treatments that the sense of the rotation of the plane of polarization is as follows. If we suppose right-hand spirals of strong bonds to be present in parallel array in the crystal, then, in general, the crystal will be laevo-rotatory, i.e. looking towards the source of light the plane of polarization is rotated anticlockwise. Such a result would be expected in the extension of the elementary concepts which were found to be applicable to the relation between double refraction and crystal structure (Wooster 1931). From the study of layer structures, e.g. $CdCl_2$, and chain structures, e.g. HgS (cinnabar), it was found that when the electric vector of a plane polarized beam vibrated parallel to the densely packed plane of atoms in a layer structure or the densely packed rows of atoms in a chain structure, the refractive index had its maximum value. Extending this relation to circularly polarized light and to spiral chains

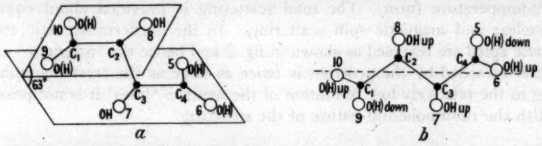


Fig. 4. (*a*) Diagram showing the form of the tartrate molecule $(CH(OH)COOH)_2$—hydrogens not attached to oxygens are omitted from the drawing. (*b*) Plan of the tartrate molecule projected on to the plane containing the four carbon atoms.

of atoms we might expect that, when the spirals representing the electric field along the light beam and the atomic spirals wind in the same sense, the interaction between the atoms and the light would be greater and the light-velocity slower. This is the result derived from the various theories. This result was derived for β-quartz by Hylleraas (1927) but the general conclusion stated above was not accepted by all scientists interested in optical activity. It was therefore most valuable to have experimental confirmation of this relationship between optical activity and structural arrangement. This was obtained by Bijvoet, Peerdeman and Bommel (1951 a, b) who determined the intensity of x-ray reflections from sodium rubidium tartrate using Zr Kα radiation. This radiation just excites the characteristic radiation of rubidium and there is consequently a difference in the intensities of the reflections from planes *hkl* and $\overline{hkl}$. The structure of the tartrate group was already known (Beevers and Hughes 1941) and hence it was possible to find the exact atomic configuration corresponding to a given handedness of the optical activity. The original convention of Emil Fischer was found to be correct. In fig. 4(*a*), the tartrate group is shown in relation to two planes

inclined at 63° to one another. In fig. 4(*b*) the same group is shown with the four carbon atoms $C_1C_2C_3C_4$ lying in the plane of the paper and the other atoms orthogonally projected on to this plane. Most of the optical polarization is associated with the oxygen atoms and we need only consider the action of atoms 5–10 on a light wave travelling from right to left. The change of direction of polarization of the OH group in going from 7 to 8 is in the sense of a right-hand spiral. Also the same is true in going from the carboxyl pair 5, 6 to the corresponding pair 9, 10. From the elementary considerations advanced above such a molecule might be expected to be laevo-rotatory and that is found experimentally to be the case.

Ramachandran (1951 b) has applied his theory to the calculation of the optical activity of $NaClO_3$ and $NaBrO_3$. He shows that the effect of all ions except the oxygens can be neglected and that the O_3 group of the chlorate radical can be replaced by a single group having two polarizabilities corresponding to vibrations in or perpendicular to the plane of the oxygen triangle. The 'reduced' O_3 groups are arranged in the structure on a distorted face-centred cubic lattice so that each one is surrounded by twelve others. Six of these O_3 groups are nearer to the one at the centre of this coordination polyhedron and the other six are further away, the distances being approximately in the ratio of 5:7. In fig. 5(*a*) is drawn a clinographic projection of a unit cell and in fig. 5(*b*) the same

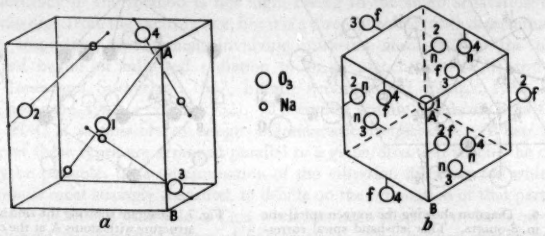


Fig. 5. (*a*) Diagram showing one unit cell of $NaClO_3$. (*b*) Projection of the unit cell of $NaClO_3$ showing the twelve neighbouring ClO_3 ions of any one such ion. Near and distant groups are denoted n and f.

cell is projected on a (111) plane. In fig. 5(*b*) the O_3 groups are represented by circles 2, 3, 4 which show their positions relative to the centres of the cube edges. The displacement along the ⟨111⟩ directions from the edge-centres is either slightly up or slightly down. The letters 'n', 'f', standing for 'near' and 'far' respectively, indicate their relative distances from the O_3 group 1 which is just below the centre of the cube in fig. 5(*b*). It will be seen that all close neighbours of group 1 are twisted clockwise relative to the face-centred-cubic positions at the centres of the edges. Also all the distant neighbours are twisted anticlockwise relative to the face-centred-cubic positions. The resultant calculated effect of this arrangement is laevo-rotatory and each unit cell behaves as a right-hand spiral of atoms. Ramachandran calculates the specific rotatory power to be 30·0°/cm for NaD which is close to the observed value of 31·3°. For a wavelength of 2573 Å the calculated value is 226° and the experimental 147·3°. This disagreement is partly due to lack of data about molar refraction for this substance.

in the ultra-violet. For $NaBrO_3$ corresponding values are: at $\lambda = 7188$ Å, 22·1° (calc), 13·9° (obs), at $\lambda = 4047$ Å, 71·0° (calc), 72·0° (obs).

A new phenomenon concerned with optical activity has been studied by Chandrasekharan (1950a). A beam of light is made to traverse a block of sodium chlorate. The slit of a spectrometer recording the scattered light is focused on the whole length of the light beam in the crystal. The incident light is plane polarized and it is rotated owing to the nature of the crystal. The scattering from any point, however, depends on the plane of polarization of the light at that point and varies as the plane rotates along the path in the crystal. The Raman lines thus show fluctuations of intensity along their lengths. From the relative positions of the maxima in the intensity of the Raman lines it is possible to assign certain frequencies to certain modes of atomic vibration and thus assist in the interpretation of the Raman lines.

The optical activity of β-quartz has been calculated from the known structure by Ramachandran (1951c). The calculated value for $\rho_{\parallel}$ the optical activity parallel to the axis is $-236°$/cm and for the direction perpendicular to this $+115°$/cm. The observed value for $\rho_{\parallel}$ is $-252°$/cm. The value of $\rho_{\perp}$ has not been measured for β-quartz but its value for α-quartz is $+126°$/cm. This value is likely to be within 5% of the value for β-quartz. In this structure (fig. 6) the

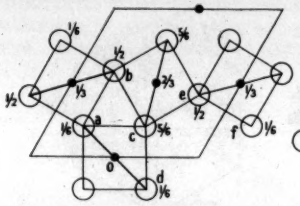


Fig. 6. Diagram showing the oxygen spiral abc in β-quartz. This left-hand spiral corresponds to a dextro-rotatory crystal.

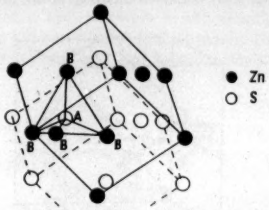


Fig. 7. Diagram showing the zinc blende structure with atoms A at the centre of a tetrahedron of atoms B.

oxygen atoms are arranged on a left-hand trigonal spiral a b c, and on a hexagonal spiral of twice the pitch, d c e f, which is right-handed. It is to be expected that the two trigonal spirals per unit cell will outweigh the effect of the single hexagonal spiral of twice the pitch. The left-hand atomic spiral should, on the elementary considerations developed above give a dextro-rotatory crystal. This is the result obtained by Ramachandran.

3.2. *Raman Spectra and Infra-red Absorption Spectra*

The main result of the great developments in the study of Raman spectra and infra-red absorption spectra of crystals has been to reveal the types of atomic vibration which are possible (Mathieu 1945). Group theory has been applied to determine what modes of vibration are consistent with the symmetry of the crystal and to predict the relationship between the Raman lines and the corresponding infra-red absorption frequencies (Bhagavantam and Venkatarayudu 1948). From the study of a number of crystals having relatively simple structures it has been possible to assign some Raman lines to oscillations of the lattice and

others to the vibrations within groups such as NH_4, CO_3, SO_4 (Couture 1944, 1948, Kastler 1948, Newman and Halford 1950, Wagner and Hornig 1950 a, b, Coté and Thompson 1951 a, b, Reding and Hornig 1951). Molecular crystals have been studied from the same point of view (Barchewitz 1943, Halford 1946, Halford and Schaeffer 1946, Carpenter and Halford 1947, Hornig 1948, 1950, Keller and Halford 1949, Karplus and Halford 1950). The Raman lines of *d*-tartaric acid, benzene in the form of a single crystal and various aromatic single crystals have all been analysed and the vibrations and translations occurring in the molecules identified (Frühling 1948, 1951, Raskin and Skripov 1949, Mair and Hornig 1949, Barer, Cole and Thompson 1949).

Though not strictly Raman lines, the Brillouin lines always accompany the Raman lines and with the aid of a high resolution spectrometer they have been used to determine the elastic properties of crystals. The Brillouin lines are due to the scattering of the light from the thermally excited elastic waves travelling through the crystal. By the well-known Doppler effect the frequency of the scattered light is slightly different from that of the incident light (Grosse 1940). This difference in frequency usually corresponds to a wave-number difference of less than 10 and is consequently difficult to observe with accuracy. This method has been applied to quartz (Krishnan and Chandrasekharan 1950), to diamond (Krishnan 1947 b, Chandrasekharan 1950 b) and corundum (Krishnan 1947 a). The accuracy of the method is not high, owing to the small separation of the Brillouin lines from the exciting line, but it is a direct check of other determinations.

An important development involving infra-red absorption is the use of polarized beams of infra-red radiation to study pleochroism (Ellis and Bath 1939, Jones and Sutherland 1947, Elliott, Ambrose and Temple 1948, Mann and Thompson 1948, Francel 1952). To certain atomic bonds such as C=O, N–H, H–O it is possible to assign characteristic frequencies. When bonds of one of these types are arranged parallel to a given direction within the crystal it may be possible, by a determination of the vibration direction for which the radiation is most strongly absorbed, to decide on the orientation of that particular type of bond in the molecule. This work has been carried out both on single crystals and on orientated layers of fine crystals. This method can be used as a refinement on the earlier study of Raman spectra. Now that the characteristic vibration frequencies for many types of interatomic bonds are well established, the pleochroism of the lines associated with these frequencies can be used to determine their orientation.

§4. PIEZOELECTRIC PROPERTIES AND ATOMIC ARRANGEMENTS IN CRYSTALS

There are two categories of piezoelectric crystal. One is characterized by a relatively small piezoelectric modulus which does not vary markedly with temperature. An example of such a crystal is quartz. The other group of crystals has similarities with ferromagnetic materials in that between certain temperatures the piezoelectric modulus is very much greater than it is outside this range of temperature. These temperatures are called Curie points, and the crystals are called Seignette-electric after Seignette salt, $NaKC_4H_4O_6.2H_2O$, for which the occurrence of very high piezoelectric moduli and hysteresis between -16°c and 24°c were first demonstrated. A rigorous study of the atomic displacements associated with the development of piezoelectric charges is not

at present possible but some progress is possible in tracing, in a qualitative manner, the relation between the orientation of polar groups of atoms and the charge development. This is what is attempted in the succeeding sections and although the phenomena are undoubtedly due to complex atomic displacements it seems worth while to try to relate the main features of the structure to the main piezoelectric effects. The crystals discussed here illustrate (i) the normal type of non-Seignette-electric crystal and (ii) the Seignette-electric type.

4.1. *Normal Piezoelectric Crystals*

We shall discuss three examples of this normal group of crystals, namely zinc blende ZnS, quartz SiO_2, and tourmaline, a complex borosilicate. These are chosen because of the simple but different features of their structures which serve to illustrate several aspects of the relation between the structure and the charge development.

Zinc blende ZnS. Zinc blende is cubic and belongs to the class $\bar{4}3m$. The symmetry elements give rise to a piezoelectric modulus d_{14} and the maximum piezoelectric moment is developed on a plate of the material cut parallel to (111), when compressed normal to the plate. The zinc atoms are distributed on a face-centred lattice and the sulphur atoms on an identical and similarly orientated lattice, displaced relative to the zinc lattice along the [111] direction through one-quarter of the length of the body diagonal (fig. 7). Each atom A is surrounded by four atoms of the other kind B distributed at the corners of a regular tetrahedron and equidistant from it. So long as the crystal is unstressed the four bonds linking any atom to its neighbours would be expected to be equivalent and no polarization of the crystal would be present. Under stress, however, this equivalence need no longer exist, for instance if a (111) plate is compressed normal to its surface the bonds parallel to the applied stress need not be affected in the same way as the other bonds. For our present purpose it is sufficient to note that any difference between the polarizations of the bonds parallel and not parallel to the applied stress will produce an electric moment, i.e. a piezoelectric effect. In zinc blende this effect might be associated with the bodily movement of the zinc lattice relative to the sulphur lattice or with a change in the polarization associated with the bonds. Either or both of these effects can lead to a development of electric moment since the crystal lacks a centre of symmetry.

α-Quartz SiO_2. α-quartz is rhombohedral and belongs to class 32. The symmetry elements give rise to two piezoelectric moduli, namely d_{11} and d_{14}. The modulus d_{11} corresponds to the production of an electric moment parallel to the axis X_1 as a result of a compression along that axis. The crystal structure may be regarded as being composed of $[SiO_4]$ tetrahedra linked at each corner to a neighbouring tetrahedron (fig. 8). The tetrahedra are orientated in three ways with respect to the diad axis OQ. The tetrahedron with its centre at level 1/3 has two sides perpendicular to the axis OQ. The tetrahedra at levels 0 and 2/3 each have one face almost perpendicular to OQ and the apexes of both tetrahedra point in directions nearly parallel to QO. It would be expected that compressing the symmetrically placed tetrahedron at level 1/3 along the direction OQ would not produce an electric moment whereas the other tetrahedra might produce an electric moment directed in the same sense for both of them.

In fig. 8 there is shown a left-hand spiral of atoms, namely a b c d e f, and, as is shown in §3, this drawing represents the atomic distribution for an optically

right-handed crystal. There is also a right-hand spiral of atoms present with its axis parallel to the optic axis but it has a repeat distance of twice the length of the one corresponding to the atoms a b c d e f and will therefore be less effective in impressing its effect on the light. This relation between the atomic distribution and the sense of the optical activity makes it possible to establish a relation between the polarity of the piezoelectric axes and the atomic arrangement. It has long been known that on a quartz crystal showing the characteristic s and x faces, fig. 9, the sense of optical activity is the same as that of the screw motion obtained by starting from the top of the crystal and travelling down over the s, x, m faces in turn (right-handed in fig. 9). It is also well known that on stretching parallel to the diad axes in turn the charges developed at the ends of these axes are as shown, i.e. negative on Q, positive on P etc. (Voigt 1910). Since the intensity of x-ray reflection from the plane $(10\bar{1}1)$, r, is greater than that from the smaller rhombohedron $(10\bar{1}1)$, z, we can determine which plane in the structure is $10\bar{1}1$ and which $\bar{1}011$. This fixes the crystallographically positive direction as OQ rather than QO, in the model of fig. 8. It was earlier suggested that the

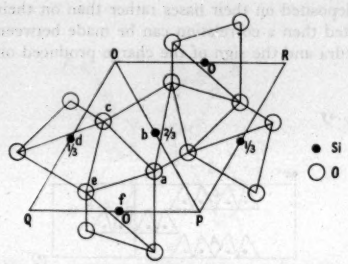


Fig. 8. Projection of the crystal structure of α-quartz on (0001).

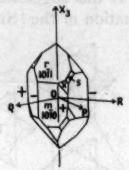


Fig. 9. Clinographic drawing of quartz crystal of right-handed optical character. The signs of the electric charge liberated on stretching parallel to OQ etc. are shown on the vertical prism edges.

main contribution to the piezoelectric effect on axis OQ was due to the deformation of tetrahedra at levels 0 and 2/3. If this is so then we obtain the result that '*on stretching the* $[SiO_4]$ *tetrahedron along a line running from one corner to the centre of the opposite face, the corner becomes positively charged and the opposite face negatively charged*'.

A series of papers (Saksena 1948, Saksena and Srivastava 1948 a, b, c) has appeared relating the piezoelectric and other physical properties of quartz to its structure. The assumption made in most of this work is that the crystal can be treated as a continuum in which the atoms move to positions determined by a homogeneous deformation of the structure. The forces involved and the charges developed are calculated from force constants derived from Raman frequencies, and certain assumptions about the distribution of charges. The method is certainly not free from objections but it is possible by means of it to calculate approximately the piezoelectric moduli from a few atomic constants.

Tourmaline. Tourmaline, which is remarkable for the variety of elements which can be found in crystals of this species, has had its structure recently determined (Donnay and Buerger 1950, Ito and Sadanaga 1951). The formula $(Na\,.\,Ca)(LiAl)_3Al_6(OH)_4(BO_3)_3Si_6O_{18}$ applies to the projections of the structure

given in fig. 10. The polyhedra contained in the structure are $[SiO_4]$ tetrahedra grouped in rings of six with their bases all parallel or nearly parallel to (0001) and their apexes all pointing in the same direction, $[BO_3]$ equilateral triangles also all parallel to (0001), $[AlO_5OH]$ or $[Al, LiO_4(OH)_2]$ octahedra, and non-regular nine-point polyhedra round Na and Ca ions. The vertical trigonal axis is polar so that on compression along this axis one end becomes positive (this is known as the *antilogous* pole) and the other (known as the *analogous*) becomes negative. The writer has observed that all tourmaline crystals which grow in radiating clusters have the analogous pole outwards from the centre of the cluster. Many other tourmaline crystals which grow in a definite direction outwards from a matrix have the same property but about one third of these rather squat crystals have the antilogous pole outwards. This may be advanced as evidence that the rates of growth at the analogous and antilogous poles are different and that the preferred growth direction is with the analogous pole outwards. It seems probable that tetrahedra of silicon atoms with four O or OH groups attached are present near the growing surface and that the rings of tetrahedra are laid down more readily when they are deposited on their bases rather than on their apexes. If this suggestion is accepted then a correlation can be made between the orientation of the $[SiO_4]$ tetrahedra and the sign of the charge produced on

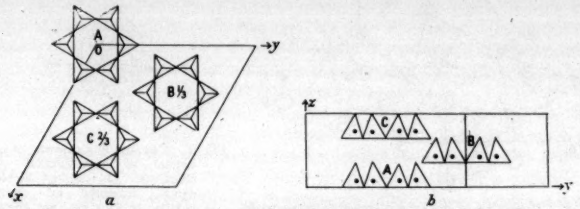


Fig. 10. (*a*) Projection of Si_6O_{18} rings in tourmaline on (0001). (*b*) Projection of Si_6O_{18} rings in tourmaline on (10$\bar{1}$0).

stretching the crystal. As already mentioned the analogous pole is charged positively on stretching the crystal along the triad axis. If the main contribution to this charge development comes from the $[SiO_4]$ group then we see that on stretching the tetrahedron along a line joining an apex to the centre of the opposite base the apex becomes positively charged. This is the same result as we derived for quartz.

4.2. *Seignette-Electric Group*

Rochelle salt, $NaKC_4H_4O_6.2H_2O$. Rochelle salt or Seignette salt shows remarkable dielectric and piezoelectric properties at room temperatures and it is not surprising that it was the first crystal of this sort to be intensively studied (Cady 1946). The crystal symmetry is that of class 222, except that between about $-16°$ c and $24°$ c there is a minute departure from orthorhombic symmetry to monoclinic, class 2 (Mueller 1935, 1940 a, b, c). The *a* axis of the orthorhombic crystal becomes the unique diad axis of the monoclinic crystal. For an orthorhombic crystal of class 222 the three relevant piezoelectric moduli are d_{14}, d_{25} and d_{36}. Of these d_{25} and d_{36} have normal values in Rochelle salt and show no remarkable change with temperature. Modulus d_{14} however is anomalous.

Below $-16°$ C it is small but at this temperature it rises at first gradually and then rapidly up to about 3×10^{-4} which is roughly 10^4 times as large as the principal modulus for α-quartz. Beyond 22·5° C the modulus d_{14} falls rapidly until at 30° C the conductivity becomes too great for further measurements.

The crystal structure (Beevers and Hughes 1941) has a remarkable feature which is brought out in fig. 11. In this projection on (001) there are included the tartrate molecules and water molecules but not the sodium or potassium ions. On the basis of the measured distances between the oxygen atoms it is concluded that bonds exist between the carboxyl oxygens and the water molecules as indicated in fig. 11 (*a*). The positions of the crosses represent conjectures as to the location of the hydrogen atoms. It will be seen that the dipoles created by this disposition of positive hydrogens all contribute to a polarization in the sense of the arrow P. In fig. 11 (*b*) the hydrogen atoms have been moved along the O–O bonds to alternative possible positions where the polarization created is just opposite to that of the previous arrangement. It should be noted that neither disposition of hydrogen atoms has orthorhombic symmetry though the

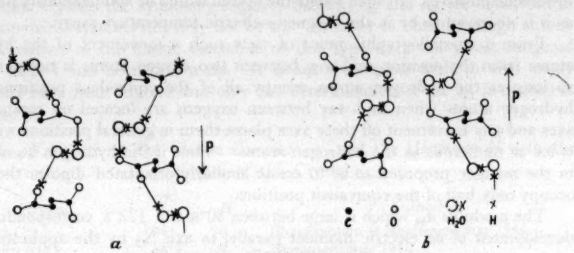


Fig. 11. Projection on (001) of chains in Rochelle salt showing the relation between direction of polarization P and the displacement of hydrogens.

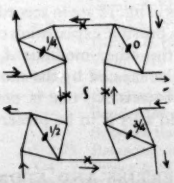


Fig. 12. Projection of PO_4 tetrahedra in KH_2PO_4 on (001). The arrows show the displacement of the crosses, representing hydrogen, if the top of each tetrahedron is negative.

rest of the atoms have this symmetry. This is not inconsistent with the crystallographic data because in the relevant range of temperature the symmetry departs definitely, though by a very small amount, from orthorhombic. It will also be seen from fig. 11 how such a disposition of hydrogen atoms could give rise to spontaneous polarization. By raising the temperature or by substituting larger ions such as NH_4 or Rb for the K, the characteristic Seignette-electric properties are lost and this may well be due to the breaking of the bonds between the oxygen atoms in the chains due to thermal agitation or distortion by the larger ions.

Potassium dihydrogen phosphate, KH_2PO_4. Potassium dihydrogen phosphate belongs to class 42m and has two piezoelectric moduli, namely d_{14} and d_{36}. Of these d_{14} is normal in value and character but d_{36} is Seignette-electric having a value of about 3×10^{-4} between the Curie points which occur at 122°K and 80°K respectively. The crystal structure (West 1930) contains $[PO_4]$ tetrahedra linked together by hydrogen bonds to form spirals parallel to the tetrad axis, X_3. In fig. 12, which is a projection on (001), these tetrahedra are represented as squares with the upper diagonal drawn in. The fractions denote the heights of the phosphorus atoms which are at the centres of the tetrahedra. The positions of the hydrogen atoms cannot be directly observed by x-rays but it is inferred that they are mid-way between oxygens of neighbouring tetrahedra as shown by the crosses in fig. 12. The system described above is electrically neutral but if the hydrogen atoms which are joined to the top of each PO_4 tetrahedron move away from it, then they necessarily move towards the bottom corners of the neighbouring tetrahedra as shown by the arrows in fig. 12. Thus the upper side of each tetrahedron becomes electrically negative and the lower side positive. In such a state the crystal would be spontaneously polarized, as it is observed to be in the Seignette-electric temperature range.

From a crystallographic point of view such a movement of the hydrogen atoms from the position mid-way between two oxygen atoms is not permitted, so long as the hydrogen atoms occupy all of the equivalent positions. The hydrogen atoms when mid-way between oxygens are located on rotation diad axes and any movement off these axes places them in general positions which are twice as numerous as the hydrogen atoms. Thus if the hydrogen atoms move in the manner proposed so as to create similarly orientated dipoles they must occupy only half of the equivalent positions.

The modulus d_{36} which is large between 80°K and 122°K corresponds to the development of an electric moment parallel to axis X_3 by the application of a torque tending to change the angle $\widehat{X_1OX_2}$. Such a torque has the symmetry of a diad axis parallel to X_3. It will be seen that the movements of the hydrogen atoms indicated by the arrows in fig. 12 are in accord with such an axis of symmetry. Thus the structure provides a possible explanation of the large value of the modulus d_{36}. It also suggests why the other modulus d_{14} is not large. Whereas it is possible to have all the dipoles created by the movement of the hydrogen atoms similarly orientated relative to axis X_3, that is not possible parallel to X_1 if the symmetry of the vertical diad axes is to be preserved.

§5. ELASTIC PROPERTIES AND ATOMIC ARRANGEMENTS IN CRYSTALS

Until the last few years it would not have been possible to write a section of this report dealing with the relation between elastic properties and the atomic arrangement because too few experimental data were available. However, it is now possible to make a rough survey of this relation. There are two ways of expressing the elastic properties of crystals, one in terms of elastic constants, usually denoted c_{ik}, where i and k may have the values 1 to 6, and the other in terms of the elastic moduli, s_{ik}. To obtain a physical meaning of the constants c_{ik} we suppose the crystal to be compressed or stretched in various directions so that only one component of strain is produced, e.g. a uniform expansion parallel to the axis X_1. Usually to obtain such a change of shape it is necessary to apply

simultaneously stretching forces of unequal amount parallel to all three axes. The constant c_{11} is the ratio under these conditions of the stretching force parallel to the axis X_1 to the expansion in that direction. We may proceed in another way to deform the crystal. A single component of stress may be applied, e.g. a stretching force parallel to the X_1 axis. Usually this will be accompanied by an expansion in the direction of the applied force together with contractions in perpendicular directions. Under certain conditions the crystal may twist when subjected to a single stretching force. Thus we fix one stress component and leave the strain components to take up whatever values they please. Under these conditions the ratio of a given strain component, e.g. the expansion parallel to the X_1 axis, to the stress component, whichever may have been applied, is an elastic modulus. For our present purpose it is more appropriate to deal with the c's rather than s's since if only one type of strain is present it is easier to relate that to the atomic arrangement.

5.1. *The Relation of c_{11} to the Nature of the Atomic Binding*

In cubic crystals the symmetry requires that c_{11}, c_{22} and c_{33} are equal but in less symmetrical systems this may not be so. In order to obtain a rough survey of the variation of the elastic constants of this type over the whole range of crystals table 1 has been drawn up. It is not a complete table in that many substances have been omitted and only one determination of the c_{11} is quoted in each case.

Table 1. Variation of the Elastic Constant c_{11} and the Hardness H with the Nature of the Crystal

Crystal	c_{11}*	H	Authority
Diamond C	94	10	Bhagavantam & Bhimasenachar 1944
Corundum Al_2O_3	46·5	9	Bhimasenachar 1950
Iron Pyrites FeS_2	36·8	6·3	Voigt 1910
Topaz AlF_2SiO_4	31·4	8	Voigt 1910
Cobalt	27		Honda & Shirakawa 1949
Beryl $Be_3Al_2(SiO_3)_6$	26	7·5–8	Liebisch 1891
Nickel Ni	25		Honda & Shirakawa 1949
Tourmaline	24	7–7·5	Voigt 1910
Haematite Fe_2O_3	24	5·5–6·5	Voigt 1910
Fluorspar CaF_2	17	4	Bergmann 1938
Germanium Ge	13		Bond, Mason, McSkimin, Olsen & Teal 1950
Lithium fluoride LiF	12		Sundara Rao 1949
Galena PbS	10	2·5–2·75	Ramachandran & Wooster 1951
Zinc blende ZnS	10	3·5–4	Prince & Wooster 1951
Barytes $BaSO_4$	9·3	2·5–3·5	Voigt 1910
Quartz SiO_2	9·0	7	Voigt 1910
Barium nitrate $Ba(NO_3)_2$	5·9		Bhimasenachar & Seshagiri Rao 1950
Rochelle salt $NaKC_4H_4O_6.2H_2O$	5·2		Sundara Rao 1949
Sodium chlorate $NaClO_3$	4·9		Ramachandran & Wooster 1951
Rock crystal NaCl	4·8	2·5	Voigt 1910
Sylvite KCl	3·8	2	Voigt 1910
Sulphur S	3·1	1·5–2·5	Sundara Rao 1950 b
Hexamethylene tetramine $C_6H_{12}N_4$	1·5		Ramachandran & Wooster 1951

The values of the elastic constants are everywhere given in units of 10^{11} c.g.s.

* For non-cubic crystals an average of c_{11}, c_{22}, c_{33} has been taken.

It will be seen from this table how great is the spread in the values of the elastic constant c_{11}. There is a rough correspondence in the order of the crystals in this table and the order of their hardnesses. This is also shown graphically in fig. 13 where $\log c_{11}$ is plotted against $\log H$. It must be remembered that

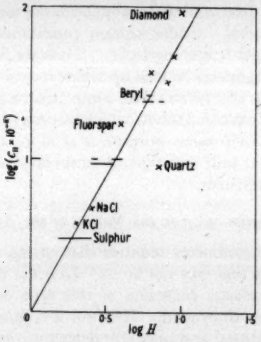


Fig. 13. Graph showing the approximately linear relation between $\log c_{11}$ and $\log H$ (H=Moh's hardness).

scratching hardness as measured on Moh's scale is not a well defined physical quantity and a strict relationship between it and any other physical property is not to be expected. The straight line drawn in fig. 13 corresponds to the relation $c_{11}=H^{7/4}$ and it will be seen that the observed points of the graph are scattered fairly evenly along this line. No great significance can be attached to this relation but it may be useful in indicating the order of magnitude to be expected for an elastic constant in a new substance.

There are two factors which are important in determining the value of c_{11}. The first is the tightness of the bonding between neighbouring atoms. The effect of this factor can be seen by comparing diamond and sulphur. In diamond strong covalent links bind each atom to its neighbours whereas in sulphur the molecules, S_8, are only held by residual bonds to one another. Corresponding to this diamond has a very high value of c_{11} and sulphur a very low value. The second factor is the rate of variation with position of the atoms of the forces of attraction and repulsion between them. An elastic constant such as c_{11} depends on the force required to produce a given percentage change in atomic distances along the axis X_1. If the law of variation of the forces of attraction and repulsion between atoms were the same in all crystals then only the absolute magnitude of these forces would affect c_{11}. This law of variation is not the same in all crystals and consequently a second factor determines the value of c_{11}. If the forces between neighbouring atoms vary more rapidly than usual, the value of c_{11} will be correspondingly greater than would be expected simply from the strength of the atomic binding. Such an effect probably explains the position of iron pyrites in the table. The general thermal and mechanical stability of this crystal is much lower than that of garnet or topaz which fall below it in the table. It is

probable that the high value of c_{11} is associated with an unusually rapid change of the interatomic forces with distance apart of the atoms.

Anisotropy in elastic properties is perhaps more marked than anisotropy in any other physical property. The first broad generalization which emerges from a survey of the elastic properties is that where the atomic binding parallel to an axial direction is strong the corresponding elastic constant is large. In table 2 the elastic moduli in the directions of strong and weak binding are given for a number of crystals.

Table 2. Variation of c_{ii} with Directions of Strong and Weak Binding

Crystal	Chemical formula	Elastic constants for directions of strong binding	Elastic constants for directions of weak binding	Authority
Cadmium	Cd	c_{11} 10·9	c_{33} 4·60	Bridgman 1924
Zinc	Zn	c_{11} 15·8	c_{33} 6·16	Bridgman 1924
Antimony	Sb	c_{11} 7·9	c_{33} 4·3	Bridgman 1924
Bismuth	Bi	c_{11} 6·3	c_{33} 4·4	Bridgman 1924
Tellurium	Te	c_{33} 7·0	c_{11} 2·9*	Bridgman 1924
Beryl	$Be_3Al_2(SiO_3)_6$	c_{11} 27·5	c_{33} 24·1	Liebisch 1891
Tourmaline	$(Na, Ca)(LiAl)_3$ $Al_6(OH)_4(BO_3)_3Si_6O_{18}$	c_{11} 27·0	c_{33} 16·1	Voigt 1910
Sulphur	S	c_{33} 4·83	$\begin{cases} c_{11} & 2{\cdot}4 \\ c_{22} & 2{\cdot}05 \end{cases}$	Sundara Rao 1950 b
Calcite	$CaCO_3$	c_{11} 13·7	c_{33} 8·0	Voigt 1910
Aragonite	$CaCO_3$	$\begin{cases} c_{11} & 16{\cdot}0 \\ c_{22} & 8{\cdot}7 \end{cases}$	c_{33} 8·5	Voigt 1910

* This value cannot be calculated from the s's exactly because s_{14} was not measured. A small error is probably caused by the arbitrary assumption that s_{14} is zero.

Cadmium, zinc, antimony and bismuth may all be regarded as layer structures in which the binding between the layers of atoms parallel to (0001) is weaker than the binding in those layers. Corresponding to this we find that c_{11} which is a measure of the resistance to expansion in the layers is greater than c_{33} which corresponds to separating the layers. Tellurium is a chain structure in which the bonds joining nearest neighbours run in vertical spirals parallel to X_3. We find, as is to be expected, that c_{33} is greater than c_{11}. Beryl and tourmaline are two ring silicates in which the Si_6O_{18} rings are perpendicular to the axis X_3. In tourmaline (Donnay and Buerger 1950, Ito and Sadanaga 1951) there are also BO_3 groups all parallel to one another and also perpendicular to the trigonal axis. Thus the relatively rigid elements of the structure are all in horizontal planes and we find as we expect that c_{11} is greater than c_{33}. Sulphur is orthorhombic and the structure is made up of S_8 rings arranged with their planes parallel to X_3 but not to X_1 or X_2. Since the molecular planes are all parallel to X_3 we should expect this to correspond to the least yielding direction and we find c_{33} is greater than c_{11} or c_{22} in conformity with this expectation. Calcite is rhombohedral and aragonite orthorhombic but in both the CO_3 groups are horizontal. The relatively rigid structural elements are again parallel to the direction of greater resistance to strain since c_{11} (c_{22}) is greater than c_{33}.

A further test of this general result can be applied by considering the anisotropy of those crystals without any well marked structural anisotropy, excluding, of course, the cubic crystals.

Magnesium and cobalt are close-packed structures in which the interatomic bonds are of equal length. Corundum, haematite and topaz are practically close-packed arrangements of oxygen ions with metal ions distributed over the interstices. Magnesium sulphate heptahydrate contains the octahedral group $[Mg.6H_2O]$ and the tetrahedral group $[SO_4]$ neither of which is very anisotropic in the distribution of bond strengths. Barytes has tetrahedral SO_4 groups and Ba ions surrounded by a relatively large number of oxygen ions. On the basis

Table 3. Variation of c_{ii} with Direction in Crystal Structures having no marked difference in binding in different directions

Crystal	Formula	Values of the elastic constants			Authority
		c_{11}	c_{22}	c_{33}	
Magnesium	Mg	5·7	—	5·9	Seitz & Read 1941
Cobalt	Co	26	—	28	Honda & Shirakawa 1949
Corundum	Al_2O_3	47	—	56	Bhimasenachar 1950
Haematite	Fe_2O_3	24	—	23	Voigt 1910
Topaz	AlF_2SiO_4	29	36	30	Voigt 1910
Magnesium sulphate heptahydrate	$MgSO_4.7H_2O$	7·0	5·3	8·2	Sundara Rao 1950 a
Barytes	$BaSO_4$	9·1	8·0	10·7	Voigt 1910

of the elementary conceptions put forward here we should not expect to observe much anisotropy in the values of the constants c_{11}, c_{22} and c_{33} and, in fact, very little anisotropy is observed.

5.2. *Theories*

The elasticity of crystals has provided subject matter for many theoretical investigations into the relation between the elastic constants and the atomic binding forces. Most success has always been achieved in connection with the alkali halides (Wasastjerna 1922, Joffe 1928, Ewald, Pöschl and Prandtl 1932, Seitz 1940, Löwdin 1948). Even the earliest theories gave a reasonably good explanation of the observed properties and subsequent theories have refined and improved this explanation. After the alkali halides the various metallic and alloy structures have been given most attention (Mott and Jones 1936, Barrett 1943, Seitz 1943, Hume-Rothery 1947). Reasonably good agreement has been obtained between calculation experiment for a number of pure metals but perhaps the most valuable contribution has come from the study of simple alloys (Jones 1934 a, b). It has now become possible, at least in certain cases, to assign contributions to the total elastic properties which are due to the various types of interatomic binding forces. It seems probable that the study of elastic properties may become of considerable importance in determining the extent to which the Brillouin zones are filled. These studies are nearly all restricted to cubic crystals because the theories become too complicated when the symmetry is lower than cubic. An exception must be made to this statement as certain calculations on quartz have been made assuming central forces are operative between the atoms and that the force constants governing their relative positions are the same as those derived from Raman spectra (Saksena 1948, Saksena and Srivastava 1948 a, b, c). This theory may be highly artificial but it seems to apply satisfactorily in some cases and although it may not be in conformity with fundamental theory of interatomic forces it might provide a valuable approximate method of calculating the elastic constants from the atomic arrangement in certain types of structure.

REFERENCES

BAGGULEY, D. M. S., and GRIFFITHS, J. H. E., 1948, *Nature, Lond.*, **162**, 538.

BANERJEE, S., 1938, *Z. Kristallogr.*, **100**, 316.

BARCHEWITZ, P., 1943, *Ann. Phys., Paris* (11), **18**, 167.

BARER, R., COLE, A. R. H., and THOMPSON, H. W., 1949, *Nature, Lond.*, **163**, 198.

BARRETT, C. S., 1943, *Structure of Metals* (New York : McGraw-Hill).

BEEVERS, C. A., and HUGHES, W., 1941, *Proc. Roy. Soc.* A, **177**, 251.

BERGMANN, L., 1938, *Ultrasonics* (London : Bell).

BHAGAVANTAM, S., and BHIMASENACHAR, J., 1944, *Nature, Lond.*, **154**, 546.

BHAGAVANTAM, S., and VENKATARAYUDU, T., 1948, *Theory of Groups and its Applications to Physical Problems* (Andhra University : Waltair).

BHIMASENACHAR, J., 1950, *Proc. Nat. Inst. Sci., India*, **16**, 241.

BHIMASENACHAR, J., and SESHAGIRI RAO, T., 1950, *Proc. Nat. Inst. Sci., India*, **16**, 235.

BIJVOET, J. M., PEERDEMAN, A. F., and VAN BOMMEL, A. J., 1951 a, *Proc. K. Akad. Wet. Ned.*, **54B**, 16; 1951 b, *Nature, Lond.*, **168**, 271.

BICKFORD, L. R., Jr., 1950, *Phys. Rev.*, **78**, 449.

BOND, W. L., MASON, W. P., MCSKIMIN, H. J., OLSEN, K. M., and TEAL, G. K., 1950, *Phys. Rev.*, **78**, 176.

BORN, M., 1933, *Optik* (Berlin : Springer Verlag).

BRIDGMAN, P. W., 1924, *Proc. Nat. Acad. Sci., Wash.*, **10**, 411.

CADY, W. G., 1946, *Piezoelectricity* (New York : McGraw-Hill).

CARPENTER, G. B., and HALFORD, R. S., 1947, *J. Chem. Phys.*, **15**, 99.

CHANDRASEKHARAN, V., 1950 a, *Proc. Ind. Acad. Sci.* A, **32**, 374; 1950 b, *Ibid.*, **32**, 379.

COTÉ, G. L., and THOMPSON, H. W., 1951 a, *Proc. Roy. Soc.* A, **210**, 206; 1951 b, *Ibid.*, **210**, 217.

COUTURE, L., 1944, *C.R. Acad. Sci., Paris*, **218**, 644; 1948, *J. Phys. Radium*, **9**, 84.

CUMMEROW, R. L., and HALLIDAY, D., 1946, *Phys. Rev.*, **70**, 433.

DONNAY, G., and BUERGER, M. J., 1950, *Acta Crystallogr.*, **3**, 379.

ELLIOTT, A., AMBROSE, E. J., and TEMPLE, R. B., 1948, *J. Chem. Phys.*, **16**, 877.

ELLIS, J. W., and BATH, J., 1939, *Phys. Rev.*, **55**, 1098.

EWALD, P. P., PÖSCHL, TH., and PRANDTL, L., 1932, *The Physics of Solids and Fluids* (London : Blackie).

FRANCEL, R. J., 1952, *J. Amer. Chem. Soc.*, **74**, 1265.

FRENKEL, J., 1945, *J. Phys., U.S.S.R.*, **9**, 299.

FRESNEL, A., 1825, *Ann. Chim. Phys.*, (2), **28**, 147.

FRÜHLING, A., 1948, *J. Phys. Radium* (8), **9**, 88; 1951, *Ann. Phys., Paris* (12), **6**, 401.

GORTER, C. J., 1936 a, *Physica*, **3**, 503; 1936 b, *Ibid.*, **3**, 1006.

GROSSE, E., 1940, *C.R. Dokl. Acad. Sci., U.S.S.R.*, **26**, 757.

HALFORD, R. S., 1946, *J. Chem. Phys.*, **14**, 8.

HALFORD, R. S., and SCHAEFFER, O. A., 1946, *J. Chem. Phys.*, **14**, 141.

HONDA, K., and SHIRAKAWA, Y., 1949, *Sci. Rep. Res. Inst. Tohoku Univ.*, **1**, 9.

HORNIG, D. F., 1948, *J. Chem. Phys.*, **16**, 1063; 1950, *Discussions of the Faraday Society*, **9**, 115.

HUME-ROTHERY, W., 1947, *Atomic Theory* (London : Institute of Metals).

HYLLERAAS, E. A., 1927, *Z. Phys.*, **44**, 871.

ITO, T., and SADANAGA, R., 1951, *Acta Crystallogr.*, **4**, 385.

JOFFE, A. F., 1928, *The Physics of Crystals* (New York : McGraw-Hill).

JONES, A. V., and SUTHERLAND, G. B. B. M., 1947, *Nature, Lond.*, **160**, 567.

JONES, H., 1934 a, *Proc. Roy. Soc.* A, **144**, 225; 1934 b, *Ibid.*, **147**, 396.

KASTLER, A., 1948, *Proc. Ind. Acad. Sci.*, **28**, 349.

KARPLUS, R., and HALFORD, R. S., 1950, *J. Chem. Phys.*, **18**, 910.

KELLER, W. E., and HALFORD, R. S., 1949, *J. Chem. Phys.*, **17**, 26.

KIRKWOOD, J. G., 1937, *J. Chem. Phys.*, **5**, 479.

KRISHNAN, K. S., and BANERJEE, S., 1935 a, *Phil. Trans. Roy. Soc.* A, **234**, 265; 1935 b, *Z. Kristallogr.*, **91**, 173; 1936, *Phil. Trans. Roy. Soc.* A, **235**, 343.

KRISHNAN, K. S., CHAKRAVORTY, N. C., and BANERJEE, S., 1933, *Phil. Trans. Roy. Soc.* A, **232**, 99.

KRISHNAN, K. S., GUHA, B. C., and BANERJEE, S., 1933, *Phil. Trans. Roy. Soc.* A, **231**, 235.

KRISHNAN, K. S., and MOOKHERJI, A., 1936, *Phys. Rev.*, **50**, 860; 1937 a, *Ibid.*, **51**, 528; 1937 b, *Ibid.*, **51**, 774; 1937 c, *Nature, Lond.*, **140**, 896; 1938 a, *Phys. Rev.*, **54**, 533; 1938 b, *Ibid.*, **54**, 841; 1938 c, *Phil. Trans. Roy. Soc.* A, **237**, 135.
KRISHNAN, K. S., MOOKERJI, A., and BOSE, A., 1939, *Phil. Trans. Roy. Soc.* A, **238**, 125.
KRISHNAN, R. S., 1947 a, *Proc. Ind. Acad. Sci.* A, **26**, 399; 1947 b, *Ibid.*, **26**, 450.
KRISHNAN, R. S., and CHANDRASEKHARAN, V., 1950, *Proc. Ind. Acad. Sci.* A, **31**, 427.
LÉVY, M., 1950, *Ann. Phys., Paris*, **5**, 153.
LIEBISCH, TH., 1891, *Physikalische Kristallographie* (Leipzig : von Veit).
LONSDALE, K., 1938, *Rep. Prog. Phys.* (London : Physical Society), **4**, 368; 1939, *Proc. Roy. Soc.* A, **171**, 541.
LONSDALE, K., and KRISHNAN, K. S., 1936, *Proc. Roy. Soc.* A, **156**, 597.
LÖWDIN, P-O., 1948, *A Theoretical Investigation into some Properties of Ionic Crystals* (Uppsala : Alinquist & Wiksells).
MAIR, R. D., and HORNIG, D. F., 1949, *J. Chem. Phys.*, **17**, 1236.
MANN, J., and THOMPSON, H. W., 1948, *Proc. Roy. Soc.* A, **192**, 489.
MATHIEU, J-P., 1945, *Spectres de vibration et symétrie des molécules et des cristaux* (Paris : Hermann).
MOTT, N. F., and JONES, H., 1936, *The Theory of the Properties of Metals and Alloys* (Oxford : University Press).
MUELLER, H., 1935, *Phys. Rev.*, **47**, 175; 1940 a, *Ibid.*, **57**, 829; 1940 b, *Ibid.*, **58**, 565; 1940 c, *Ibid.*, **58**, 805.
NÉEL, L., 1948, *Ann. Phys., Paris*, **3**, 137; 1950, *C. R. Acad. Sci., Paris*, **230**, 190.
NÉEL L., and BROCHET, P., 1950, *C.R. Acad. Sci., Paris*, **230**, 280.
NEWMAN, R., and HALFORD, R. S., 1950, *J. Chem. Phys.*, **18**, 1276.
PASCAL, P., 1913, *Ann. Chim.*, **29**, 218.
PAULING, L., 1936, *J. Chem. Phys.*, **4**, 673.
PRINCE, E., and WOOSTER, W. A., 1951, *Acta Crystallogr.*, **4**, 191.
RAMACHANDRAN, G. N., 1951 a, *Proc. Ind. Acad. Sci.*, **33**, 217; 1951 b, *Ibid.*, **33**, 309 ; 1951 c, *Ibid.*, **34**, 127.
RAMACHANDRAN, G. N., and WOOSTER, W. A., 1951, *Acta Crystallogr.*, **4**, 431.
RASKIN, SH. SH., and SKRIPOV, F. I., 1949, *Dokl. Akad. Nauk. SSSR.*, **64**, 317.
REDING, F. P., and HORNIG, D. F., 1951, *J. Chem. Phys.*, **19**, 594.
ROOKSBY, H. P., and TOMBS, N. C., 1951, *Nature, Lond.*, **167**, 364.
SAKSENA, B. D., 1948, *Nature, Lond.*, **161**, 283.
SAKSENA, B. D., and SRIVASTAVA, K. G., 1948 a, *Proc. Ind. Acad. Sci.*, **28**, 423; 1948 b, *Ibid.*, **28**, 437; 1948 c, *Ind. Phys.*, **22**, 475.
SEITZ, F., 1940, *The Modern Theory of Solids* (New York : McGraw-Hill); 1943, *The Physics of Metals* (New York : McGraw-Hill).
SEITZ, F., and READ, T. A., 1941, *J. Appl. Phys.*, **12**, 100.
SERRES, A., 1932, *Ann. Phys., Paris*, **17**, 5.
SHULL, C. G., and SMART, J. S., 1949, *Phys. Rev.*, **76**, 1256.
SMART, J. S. and GREENWALD, S., 1951, *Phys. Rev.*, **82**, 113.
SUNDARA RAO, R. V. G., 1949 a, *Proc. Ind. Acad. Sci.*, **30**, 173; 1949 b, *Curr. Sci.*, **18**, 436; 1950 a, *Proc. Ind. Acad. Sci.* A, **31**, 365; 1950 b, *Ibid.*, **32**, 275.
TOMBS, N. C., and ROOKSBY, H. P., 1950, *Nature, Lond.*, **165**, 442; 1951, *Acta Crystallogr.*, **4**, 474.
VOIGT, W., 1910, *Lehrbuch der Kristallphysik* (Leipzig : Teubner, 1st. edn.).
VOLKENSHTEIN, M. V., and BYUTNER, E. K., 1951, *Dokl. Akad. Nauk. SSSR.*, **76**, 575.
WAGNER, E. L., and HORNIG, D. F., 1950 a, *J. Chem. Phys.*, **18**, 296; 1950 b, *Ibid.*, **18**, 305.
WALLER, I., 1932, *Z. Phys.*, **79**, 370.
WASASTJERNA, J. A., 1922, *Z. phys. Chem.*, **101**, 193.
WEST, J., 1930, *Z. Kristallogr.*, **74**, 306.
WHEATLEY, J., and HALLIDAY, D., 1949, *Phys. Rev.*, **75**, 1412.
WOOD, W. W., FICKETT, W., and KIRKWOOD, J. G., 1952, *J. Chem. Phys.*, **20**, 561.
WOOSTER, W. A., 1931, *Z. Kristallogr.*, **80**, 495.
ZAVOISKY, E., 1945, *J. Phys., U.S.S.R.*, **10**, 170.

General Reference

Strukturbericht, Vol. I–VII, 1931 (Leipzig : Akad. Verlag).

RAMAN EFFECT IN SOLIDS

By A. C. MENZIES
Hilger and Watts, Ltd., London

CONTENTS

	PAGE
§ 1. Introduction	83
§ 2. Early work	84
§ 3. Experimental technique	84
§ 4. Theoretical considerations	85
4.1. Normal vibrations	85
4.2. Fundamental or 'homogeneous' vibrations	86
4.3. Second order Raman spectrum	86
4.4. Classification of homogeneous vibrations into 'internal' and 'external'	87
4.5. Classification of external vibrations into 'translational' and 'rotational'	87
4.6. Coupling between vibrations belonging to the same mode	88
4.7. Internal vibrations	88
4.8. Application of group theory to determine the number and types of homogeneous vibrations	88
4.9. Variations of depolarizations and intensities of Raman lines in crystals with orientation	89
§ 5. Alkali halides	89
5.1. Born–von Kármán theory	89
5.2. Simple case	90
5.3. Methods of solution	91
5.3.1. Direct solutions	91
5.3.2. Montroll's method of solution	93
5.3.3. Houston's method of solution	94
5.3.4. Criticism of the Montroll and Houston methods	95
5.4. Anharmonicity of vibrations	96
5.5. Raman's theory of lattice dynamics	97
§ 6. Alkali halides: experimental	98
§ 7. Concluding remarks	99
§ 8. References	100
8.1. Theory	100
8.2. Experimental technique	102
8.3. Inorganic solids (other than alkali halides)	103
8.4. Alkali halides (experimental)	106
8.5. Organic solids	106

ABSTRACT. A summary is given of the early work in this field, followed by an outline of the theoretical treatment of vibrations in crystals. The alkali halides have received a somewhat more detailed treatment than other substances. A bibliography is appended, giving references to many papers, covering more especially the last five years.

§ 1. INTRODUCTION

At a rough estimate several thousand papers dealing with the Raman effect have appeared since its discovery by Sir C. V. Raman in 1928, and nearly a thousand of these have been concerned with solids.

In the attempt to produce a report on progress in this subject in a reasonable compass, detailed description is confined to work done in the last five years. For the benefit of those not closely engaged in this field a summary is given of the work of the earlier period and certain terms are defined.

Some aspects of the theory are still controversial, and with regard to these various views are presented.

It is not possible to write on this subject without making many references to infra-red spectra since, depending upon the symmetry of the crystal, many vibrations may appear in both the Raman effect and the infra-red spectrum. To save space the abbreviations R.E. and I.R. will be used in the text for these two phenomena respectively.

A bibliography is appended giving many references, followed in each case by a few words indicating the main point connected with this report. The greater part of the papers mentioned have appeared in the last five years, and only a few are referred to in the text, since the object is to assist those who wish to follow up particular points. It is not claimed to be exhaustive. The references have been grouped into sections according to the main content of the paper, but it is to some extent arbitrary, since most of the papers are mixed in character.

§ 2. EARLY WORK

An account is given by Sir C. V. Raman (1928) of his first finding the Raman effect by using sunlight and complementary filters. He has told the writer how he later sat up all one night, putting one substance after another before the slit of a small Hilger quartz spectrograph, in the light of a mercury vapour lamp. This was, he added, the first time he had used the instrument. He investigated solids as well as liquids, e.g. ice and glass.

Shortly afterwards, Landsberg and Mandelstam (1928) published an account of their independent discovery, using a block of quartz as the scattering material.

Several papers followed, by various investigators, in which crystals were used, but the work was confined to those few materials which could be obtained in pieces of fair size. Menzies (1929) and, shortly afterwards, Bär (1929) independently showed that the effect might also be worked upon with powdered crystals. This greatly extended the available field, and a number of papers appeared, notably by Indian workers, on the R.E. in solids in powder form. W. Gerlach (1930) introduced a technique of complementary filters into powder work which was improved upon by Médard (1934) and Ananthakrishnan (1937). It is, however, always preferable to use a single crystal if it is available, on account of the shorter exposures resulting in practice from the smaller amount of parasitic light. It is essential to use a single crystal when polarizations of the lines are to be investigated.

§ 3. EXPERIMENTAL TECHNIQUE

As the source of the incident light the majority of workers have used the mercury vapour lamp, and most often the 4358 and 4047 Å lines, unless the material is coloured, when 5461 Å (mercury) or the D lines from a sodium lamp can be employed, according to the transmission of the material. In the ultra-violet region 2536 Å (mercury) is used for preference.

When it is important to avoid ambiguity of the exciting line, filters may be used, such as sodium nitrite solution to suppress 4047 Å, iodine dissolved in CCl_4 to suppress 4358 Å. All mercury lamps have, to greater or lesser degree, a continuum to the red side of 4358 Å, and when it is important to suppress this a filter such as praseodymium chloride solution is inserted.

The above filters are placed between the lamp and the specimen. Sometimes the exciting line itself is strong in the spectrum of the light emerging from the

specimen, especially when powdered crystals are used. This may cause much scattered light in the spectrograph, and to weaken this a filter suppressing the exciting line is placed between the specimen and the spectrograph.

Special mention should be made of the ingenious complementary filter technique devised by Rasetti (1932) in which the resonance line 2536 Å of mercury is used to excite the R.E., and is almost completely suppressed after the scattering by a selective absorbing filter of mercury vapour. This method is of limited application, due to ultra-violet absorption and to the fluorescence which may be caused by the short-wave exciting light.

As an example of the use of complementary filters with powders may be mentioned the arrangement of Ananthakrishnan. He used 4047 Å as the exciting radiation, weakening 4358 Å with a solution of iodine in CCl_4. The scattered light was passed through a solution of sodium nitrite, thus suppressing the exciting radiation. Unfortunately, the iodine solution cannot be sufficiently strong to suppress 4358 Å entirely, or 4047 Å itself is adversely affected: the result is that frequencies near to 1750 cm^{-1} may be difficult to see in the spectrum. In addition, the sodium nitrite suppresses the spectrum some way to the red side of 4047 Å, so that Raman lines corresponding to small frequency shifts are also suppressed. Another drawback is that 4047 Å can on occasion cause objectionable fluorescence. Ananthakrishnan also used a similar technique for 4358 Å as the exciting line, but it was less efficient.

Monochromators have also been used to purify the radiation incident on the specimen: they prolong the exposure time, but use of them is justified in work with powders. Two examples of progress in this direction are the apparatus of Reitz (1940), being an improvement of an earlier arrangement devised in the school of Kohlrausch, and that of Harrand (1948).

Interference filters are now available which can be used as complementary filters, using a filter transmitting the exciting line between lamp and specimen, and reflecting the scattered light on to the spectrograph slit with a similar filter so that all light other than the exciting light is reflected. This method has now been used by Güber and Riggert (1951).

Spectrographs for the purpose have usually two or three large prisms (aperture about 3 inches square) or a large grating. The cameras vary in aperture ratio from about F/6 to about F/1. Almost all the work on solids has been done with photographic technique. Recording Raman apparatus is now available commercially, but has so far hardly been used for this purpose.

§ 4. THEORETICAL CONSIDERATIONS

4.1. *Normal Vibrations*

Any general vibration may be synthesized by superposing normal vibrations having suitable respective amplitudes and phases. Each harmonic normal vibration has one, and only one, frequency connected with it. In a non-degenerate normal vibration all particles taking part in it move in straight lines in phase.

In a real degenerate normal vibration (i.e. one degenerate for reasons of symmetry, and not an accidentally degenerate one due to special values of masses and forces) which is harmonic all particles move with one frequency, but not necessarily along straight lines. If the real degeneracy is double, the particles may be thought of as moving round ellipses and, if triple, on the surfaces of ellipsoids. These degenerate normal vibrations arise when the symmetry does

not fix the motion in an unique direction, and it may take place with the same frequency in more than one direction—e.g. round an axis, or in a plane if doubly, and in any direction in space if triply degenerate.

For combinations and overtones the normal vibrations taking part may have different phases in different cells, provided that there is a fixed relationship between the phase of the movements in one cell and that in another specified cell. This leads to a continuum of overtones and combinations.

Sir C. V. Raman applies to crystals the definition of the normal vibration stated by Lord Rayleigh (1926) in his *Theory of Sound*, viz. "all particles in a normal vibration have at any instant the same or opposite phase of vibration". According to Raman this results in a great restriction of the number of vibrations classed as normal ones, so that the continuum is replaced by a finite number of sharp lines.

4.2. *Fundamental or 'Homogeneous' Vibrations*

Among the normal vibrations which may occur in a crystal there are certain ones of particular importance.

Every crystal may be considered as built up by the three translations of an elementary cell, the translational unit cell. If each such cell contains r atoms, it possesses $3r$ degrees of freedom. Normal vibrations will be possible in which the contents of every translational unit cell are moving in the same way, so that the wavelength in the crystal is infinite. If ϕ_{kl} is the phase difference for the vibration between cell k and cell l, the condition may be written $\phi_{kl}=0$. An example is the 164 cm^{-1} vibration of rocksalt, in which all Cl ions are moving simultaneously in one direction and all Na ions oppositely. Such vibrations are usually termed 'fundamental'. (This is an unfortunate choice, in view of the clash with the same word used in a different sense in group-theory—a better term would be 'homogeneous'.) Since the translations of the atoms are also included, the number of homogeneous *vibrations* is $3r-3$. In the example of NaCl, since $r=2$ there are three such vibrations, but since the motion may occur in any direction in space it is triply degenerate, and only one numerical value is associated with this vibration. (This is a real-degenerate vibration, arising from the symmetry, and it could only be split by applying an appropriate distortion which would sufficiently lower the symmetry.)

The homogeneous vibrations, when active, constitute the *first order* Raman spectrum of the crystal.

4.3. *Second Order Raman Spectrum*

When an exposure is given of considerably extended duration, a faint spectrum appears in addition to any first order spectrum which may be present also, and which is now relatively over-exposed. In the alkali halides, where there is no first order Raman spectrum, the second order spectrum is the first which appears.

The vibrations which give rise to this spectrum need no longer be homogeneous ones, although some may correspond to overtones of homogeneous vibrations, as for example in diamond, where the overtone of the 1332 cm^{-1} vibration appears in considerable strength. The new spectrum arises through anharmonicity of the vibrations, either mechanical or electrical or both. (Krishnan and Roy (1951, 1952) have shown recently that in alkali halides electrical anharmonicity alone is important.) Suppose that there are two vibrations A and B for which the phase differences between any cell k and any cell

l are ϕ^{A}_{kl} and ϕ^{B}_{kl} respectively. Then a condition for the appearance of the combination of these two vibrations is $\phi^{A}_{kl}+\phi^{B}_{kl}=0$. If A and B correspond to wave motions propagated in *opposite* directions, having the same wavelength in the crystal, then this condition can be fulfilled, and it is possible for an octave to be allowed for any wavelength in the crystal down to twice the width of the unit translational cell. Virtually all frequencies can be present, from an upper limit downwards, so that a continuum is found (if symmetry allows it). For example, in rocksalt a continuum is to be expected on this view, from $2\nu_0$ downwards, where ν_0 is the frequency of the first order homogeneous vibration.

4.4. *Classification of Homogeneous Vibrations into 'Internal' and 'External'*

While in all cases the ultimate masses moving in the vibration are atoms, either ionized or not, it is convenient to consider groups of atoms in the crystal variously referred to as 'islands', 'blocks', 'complexes'.

An 'island' is a group of atoms more closely bound to one another than they are to neighbouring islands. For example, in calcite the forces binding the atoms together in the CO_3 ion are much stronger than the forces binding that ion to any calcium ion. The result of this is that the Hamiltonian for the vibrations contains terms which may to a first approximation be separated from one another, and one may in practice recognize separately the internal vibrations, which take place between atoms of the CO_3 ions, and external vibrations, which take place between islands as units.

In a crystal such as rocksalt the atoms within one cell are no more strongly bound together than they are to atoms in other cells, and so the only homogeneous vibrations occurring are external, or lattice vibrations.

Because of the smaller forces involved, the external vibrations are commonly found around 50 to 150 cm^{-1}, while the internal vibrations are much higher, around 300 to 3000 cm^{-1}.

4.5. *Classification of External Vibrations into 'Translational' and 'Rotational'*

In an external vibration the island moves to a first approximation as a rigid body, and in the vibration may conveniently be regarded as either rotating or translating.

In the rotational oscillations the centre of gravity of the island remains fixed, and the island librates. The nature of these librations ('pivotements') has been investigated, notably by Kastler and Rousset (1941). They showed that the librations may be regarded as taking place about the principal axes of inertia of the island. The axes of the librations are often conditioned by the symmetry in the crystal.

If the symmetry permits it, these librations are active in the Raman effect, since the polarizability ellipsoid of the island librates. If the island is very anisotropic the Raman lines resulting may be very strong.

In the translational vibrations the island moves without turning, so that the vibration may be described as that of the centres of gravity of the islands. If symmetry permits it these vibrations also are active in the Raman effect, but since the activity results from the effect of different islands on each other, the more loosely bound they are together the less the effect and consequently the weaker are the Raman lines resulting. This is particularly true of organic crystals, where the islands are very loosely bound.

External frequencies are dealt with in detail in Chapter III of the monograph by Rousset (1947).

4.6. *Coupling between Vibrations belonging to the same Mode*

It should be observed that the classification of vibrations into 'internal' and 'external', and of the latter into 'rotational and 'translational' is one of convenience. There is nothing in the symmetry to justify this separation: the considerations of symmetry lead only to classifying vibrations under different modes of vibration, all vibrations belonging to one mode being subject to the same transformation law.

Kastler (1948) has drawn attention to this in connection with the division of external vibrations into translational and rotational types. (He could have gone further and included also the internal vibrations.) If two vibrations, one of rotational and one of translational type, belong to the same mode, then they are coupled, and if the nature of a vibration is such that it is strong in the R.E., the other will appear also in strength, even though an examination of the form of the vibration would not lead one to expect it. More accurately speaking, two frequencies are observed which are the result of the coupling of the two vibrations. This accounts, for example, for the observation of more active external frequencies in calcite than would be expected if the two vibrations were considered separately: one of the two would be expected to be very weak, the other strong, but the coupling makes both appear.

4.7. *Internal Vibrations*

When the crystal contains islands they may be considered in isolation, to obtain the internal vibrations. The treatment developed by Cabannes (1932), Placzek (1934) and others may be applied to the island to determine the number, degeneracies, activities and depolarizations of the vibrations to be expected of it. Experimentally, frequencies are found for the internal vibrations of the islands near to (but not identical with) those exhibited by them in the free state, e.g. the CO_3 ion in calcite and in solutions of carbonate give rise to frequencies which are nearly the same.

The difference arises from the lowering of symmetry which the free complex undergoes when it is embedded in the crystal. This may frequently be deduced from the results of x-ray studies of the crystal. The lowering of symmetry may result in vibrations previously degenerate being split, and in vibrations previously inactive becoming active. Thus, the island in the crystal in general gives rise to more Raman lines than when free. These effects have been studied notably by Couture and Mathieu (e.g. Couture 1947) and by many of the Indian workers.

4.8. *Application of Group Theory to determine the Number and Types of Homogeneous Vibrations*

A considerable simplification was made by the application by Placzek (1934) and by Tisza (1933) of group-theory to the treatment of molecular vibrations. Tables based on these methods which are applicable to determine the homogeneous vibrations of crystals have been published (e.g. in the monograph by Mathieu (1945)).

For each point symmetry group the tables list the fundamental normal nodes, and against each the degeneracy, the degrees of freedom associated with each symmetry element of the group, and the components of a vector (I.R.) and the components of a second order symmetrical tensor (R.E.) compatible with that mode.

Having inferred from x-ray evidence the symmetry group of the crystal, one deduces the point symmetry group which results when the translations are

suppressed. The atoms of the unit cell which result from the suppression of translations lie on certain symmetry elements of the point group, and the number of these for each symmetry element determines the number of degrees of freedom in each normal mode. Thus, from the tables it is possible to deduce how many vibrations are active in the R.E. (and in the I.R.), which components of the Raman tensor are involved with each, how many vibrations are internal or external and, if the latter, whether they are of rotational or translational character.

4.9. *Variations of Depolarizations and Intensities of Raman Lines in Crystals with Orientations*

The depolarization is a valuable guide in allocating observed frequencies to types of vibration. The depolarization may be deduced from the Raman tensor components connected with the mode of the vibration, once the orientation of the crystal axes relative to the axes of the experiment has been determined. Saksena (1940) gave explicit results for the depolarizations, and total intensities for the different modes of vibration in the different crystal classes, for transverse and longitudinal scattering. These were given for three orientations of the chief axis (in the non-cubic classes) and for three orientations of a cubic axis in the cubic class. Both the intensity and the depolarization of lines vary according to the orientation of the crystal during the scattering experiment, even in the case of *cubic* crystals. This point has been brought out strongly by Couture and Mathieu (1948 b). The Raman activity is connected with the rate of change of the polarizability tensor with displacement during the vibrations. Although the induced moment remains proportional to the applied field, the Raman tensor loses its cubic symmetry for all vibrations other than totally symmetric ones. For cubic crystals the homogeneous vibrations active in the R.E. are of type A (totally symmetric and non-degenerate), type E (doubly degenerate) and type F (triply degenerate).

If the experiment is arranged with light incident along $x0$, and scattered along $0y$, the cubic crystal may be arranged with its cubic axes along $0x$, $0y$, $0z$, in which case the depolarization for A and E type vibrations is zero, and for F type is 2. If, however, the crystal is cut so that the appropriate edges of the block are parallel to one threefold axis, and to two directions perpendicular to it, then the depolarizations found for E-type vibrations depend upon the direction of the threefold axis during the experiment, being 2, 1, $\frac{3}{4}$ respectively according as the threefold axis is parallel to $0x$, $0y$ or $0z$. Other changes occur for the F-type vibrations, and for other ways of orientating the crystal.

The intensity of the line corresponding to a vibration also varies markedly with the orientation.

Consequently, by making exposures with polarized light with different orientations of the crystal, the variations in intensity and depolarization indicate the mode to which the observed frequency belongs.

§ 5. ALKALI HALIDES

5.1. *Born–von Kármán Theory*

The equations for the natural frequencies of vibration of a cubic ionic crystal were stated by Born and von Kármán in simple form as long ago as 1912, but the mathematics involved in deducing therefrom the actual frequencies of a particular crystal are very complicated and, even when approximations are introduced, the numerical calculations are tedious. In their treatment Born and von Kármán

assumed quasi-elastic forces between neighbouring and next nearest particles. They wrote down the equations for a set of N coupled oscillators having the coupling and force constants of the crystal. The normal modes of vibration were then the characteristic roots of a secular determinant of order $3N$. The frequency distribution $g(\nu)$ was then to be obtained by finding the roots of all the $3N$ equations and then calculating the number of roots in each small interval of frequency.

5.2. *Simple Case*

It will perhaps assist those not familiar with this subject if brief consideration is given to the problem of the one-dimensional lattice, particularly as the solution has points of resemblance with those for more complicated cases.

Imagine an indefinitely extended linear chain of masses M and m alternately, evenly spaced. Suppose that displacements occur only in the direction of the chain, and that, between neighbouring *like* atoms, the increase of phase of the motion, ϕ, is constant. Let α be the force constant between neighbouring atoms. On writing down the equations of motion one gets the determinantal equation

$$\begin{vmatrix} M\omega^2-\alpha & 2\alpha\cos\phi/2 \\ 2\alpha\cos\phi/2 & m\omega^2-\alpha \end{vmatrix} = 0,$$

which leads to

$$\omega^2 = \left[\frac{\alpha}{\mu} \pm \left(\frac{\alpha^2}{\mu^2} - \frac{4\alpha^2\sin^2\phi/2}{Mm}\right)^{1/2}\right],$$

writing μ for the reduced mass and ω for $2\pi\nu$. This has two sets of solutions for ω, depending on whether the positive or negative sign is chosen, and, when expressed as a graph of ω against ϕ, two curves are obtained, the upper one (positive sign) referred to as the 'optical' branch, the lower one (negative sign) referred to as the 'acoustical' branch. The separation of the two branches at their mid-points depends upon the mass-ratio m/M.

The form of the curves is as shown in fig. 1. For a particular state of vibration ϕ will have a particular value, say ϕ', and there are two frequencies ω_A' and ω_B' in the optical and acoustical branches respectively. The state of vibration with $\phi=0$ gives rise to the frequencies ω_0 and zero. The motion giving rise to ω_0 is that in which the M lattice as a whole moves against the m lattice, and the frequency is that of the fundamental I.R. absorption frequency of the 'crystal'.

The state of vibration for which $\phi=\pi$ gives rise to two frequencies ω_A and ω_B. The motion corresponding to the former is that of the lighter atoms (m, say) alternately vibrating against one another while the M's are at rest, and vice versa for ω_B.

It will be seen from the figure that for this special case, the density of frequencies will have maximum values for ω_0, ω_A and ω_B.

For this case the selection rules will allow in the I.R. the frequencies ω_0, $(\omega_A' \pm \omega_B')$ to be active, that is to say, the combination of frequencies belonging to the same state of vibration, but derived from different branches, is active. In the R.E. the octaves $2\omega_A'$, $2\omega_B'$ are active, i.e. the sums of frequencies belonging to the same state of vibration and the same branch are active. As m approaches equality to M the curvatures at the mid-points of the two curves become more pronounced and the curves approach one another at the mid-points, until at $m=M$ they cross one another at the common mid-point, so that $\omega_A=\omega_B$ and the frequency-density is no longer a maximum there.

5.3. *Methods of Solution*

The solution of the Born–Kármán equation has been attempted in several ways, and the exactness of the solution is proportional to the mathematical effort expended, which is in any case considerable. The principal methods of solution have been: 5.3.1. Direct solution by considering simplifications, and numerical computation of the frequencies for a number of special values of the phase differences, so building up a distribution function which is more correct the greater the number of states for which frequencies are computed (e.g. Kellermann 1940). 5.3.2. Reducing the problem to that of the moment problem

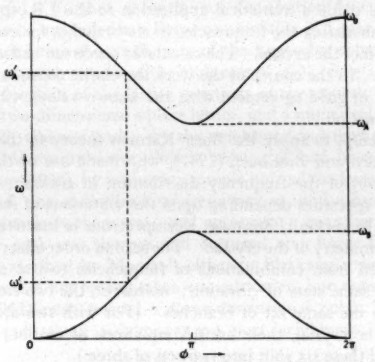


Fig. 1. Frequency–phase curves for $m/M=0{\cdot}49$ (KBr).

in statistics (Montroll 1942). 5.3.3. Calculating the distribution function along selected directions in the reciprocal lattice and interpolating between them (Houston 1948). These three methods will be considered separately They have been applied to various special cases which allowed simplification.

5.3.1. *Direct solutions.*

In a series of papers, the first one in collaboration with Born, Blackman has extended the theory and applied the results to special problems, such as specific heats. (A report on this application was published by him in *Reports on Progress in Physics*, Vol. VIII, 1938.)

Born and Blackman (1933) obtained numerical values of the subsidiary I.R. maxima for LiF, NaCl and KCl, attributing their existence to the anharmonic nature of the vibrations. Blackman (1933) generalized this treatment to the three-dimensional case and, in 1935 (1935 a, b), applied the theory to account for the divergence of observed specific heats from the predictions of the Debye theory. Blackman (1937 a) returned to the I.R. problem and investigated the one-, two-, and three-dimensional cases, examining the separation into optical and acoustical branches. Observed subsidiary I.R. maxima were attributed to summational tones of frequencies associated with density maxima in the theoretical continuous spectrum of normal vibration frequencies. These summation tones appeared because the frequencies giving rise to them were

coupled with the reststrahlen frequency. Anharmonic vibrations were assumed. The difference tones corresponding to the summational ones lay far outside the observed region. The results appeared satisfactory for NaCl but not for KCl.

Lyddane and Herzfeld (1938) investigated the three-dimensional case for coulomb forces and a short-range repulsion.

Kellermann (1940) treated the three-dimensional problem of the NaCl lattice, first as an electrostatic one, as did Lyddane and Herzfeld, and then by finding a solution of Maxwell's equations for the electromagnetic field in the crystal. The latter treatment was necessary for the case of infinitely long wavelengths in the crystal, while the purely electrostatic solution suffices for all other cases. He made a numerical application to the I.R. spectrum of NaCl. This involved calculating the frequencies in six branches for each of 48 different states of vibration of the crystal. The calculated spectrum had a fair resemblance to that observed. In the course of the work the elastic constants were computed and found to be in good agreement with the known values within the accuracy of their determination.

The first attempt to apply the Born–Kármán theory to the R.E. in crystals was made by Born and Bradburn (1947), who made use of the calculations of Kellermann (1940) of the frequency distribution in rocksalt. The first order effect was a line spectrum depending upon the vibrations of infinite wavelength in the crystal. In rocksalt, however, this spectrum is inactive in the R.E., on account of the symmetry of the crystal. The second order effect was a continuous spectrum, derived from combinations of frequencies two at a time, each pair belonging to the same state of vibration; moreover, the two frequencies in each pair belonged to the same set of branches. (For with two atoms in each cell, and N cells in the crystal, there are $6N$ equations of motion, which split into N sets of six, and these six split into two sets of three.)

In making their numerical computations Born and Bradburn calculated the eigenvectors for only one point in reciprocal space, i.e. in their notation $q=\frac{1}{2}\frac{1}{2}\frac{1}{2}$. This corresponds in the one-dimensional case of § 5.2 to the vibrational state given by $\phi=\pi$. This approximation is probably too severe to give more than qualitative value to their solution. They showed, however, that the treatment could yield the general form of the observed second order R.E. in rocksalt and similar crystals, giving the hope that a more nearly perfect approximation could provide better agreement.

A similar attack was made on the problem of the R.E. in diamond by Smith (1948), assuming arbitrary forces between neighbouring atoms and central forces between next nearest neighbours. She obtained the force constants from the known experimental values of the elastic constants and the frequency shift of the strong Raman line. The frequency distribution was then calculated and used to find the specific heat. Comparison with the known specific heat–temperature curve showed that neighbour forces alone were insufficient to secure agreement, and that next nearest neighbour forces had to be brought in to do so. This gave confidence in calculating the second order Raman spectrum, but, in order to make the work not excessively laborious, the same approximation was made as that already made by Born and Bradburn in attaching special importance to the state $q=\frac{1}{2}\frac{1}{2}\frac{1}{2}$. The resulting frequency density curve had a general resemblance to the observed second order Raman spectrum. The general result of this type of calculation has been, so far, to produce a qualitatively correct picture, but some of the peaks are less pronounced than the experimental

results show. More elaborate calculations might well show greater agreement.

In the next section is considered a method of solution which emphasizes the sharpness of certain peaks in the continuum.

5.3.2. *Montroll's method of solution.*

Montroll (1942) used the matrix theorem that the sum of the nth powers of the characteristic values of a matrix is equal to the trace of the nth power of that matrix. He saw that the sum of the nth powers of the characteristic values of the Born–Kármán determinant expressed as a matrix was proportional to the nth moment of $g(\nu)$. Further, a continuous distribution function defined in a finite interval may be expanded as a linear combination of Legendre polynomials. Hence he concluded that the coefficients of the Legendre polynomials could be expressed as linear combinations of the moments of $g(\nu)$. So $g(\nu)$ could be stated analytically as a function of the lattice constants of the crystal. This treatment was applied to a two-dimensional square lattice, and a numerical case was worked out for $\gamma/\alpha = 1/20$, where α and γ are the neighbour and next neighbour forces respectively. This example was chosen because it had been already worked upon by Blackman (1936), by evaluating sample normal modes. Two prominent peaks were found, one at the limiting frequency ν_L, and a broad one at about $\nu/\nu_L = 0{\cdot}4$. Qualitatively there was good agreement: quantitatively Blackman's lower frequency peak was sharper and occurred at a lower frequency. This discrepancy was ascribed by Montroll either to Blackman's smoothing process or to his (Montroll's) not having taken enough orders into consideration.

Montroll (1943) next treated a three-dimensional cubic lattice, assuming central forces, and taking neighbour and next neighbour interactions only. This gave

$$\frac{\nu_L g(\nu)}{3(2N)^3} = \frac{(\nu/\nu_L)^2}{2\pi^2\tau},$$

where τ was given by $2\gamma/\alpha = \tau(1-2\tau)^{-1}$ and the other symbols were as before. Plotting the left-hand side of the above equation against ν/ν_L for $\tau = 0{\cdot}075$, $0{\cdot}1$, $0{\cdot}15$, he obtained in all three cases curves showing a low-frequency peak, and a high-frequency peak (near ν_L). The height of the low-frequency peak increased, and that of the high-frequency peak decreased, with increasing τ. He interpolated for the value of $\tau = 0{\cdot}085$ used by Blackman and obtained excellent agreement with his results. The aspect of Montroll's results of greatest interest for the comparison with experiment is that the two peaks are logarithmic infinities, i.e. the expression for $g(\nu)$ contains a factor of form $\log(x-y)$ and at the two critical frequencies $x \to y$. The area under the curve, however, remains finite. Consequently the continuous spectrum at these places will assume the character of a line.

Montroll and Peaslee (1944) applied this treatment to a body-centred cubic lattice, assuming neighbour and next neighbour interactions to derive $g(\nu)$. They found two peaks, as for the simple lattice, but the two were more nearly equal. In 1947 Montroll applied his treatment to the monatomic square lattice, once more assuming neighbour and next neighbour interactions. He calculated first the special case of $\tau = \frac{1}{3}$, since this resulted in a closed expression for $g(\nu)$. The characteristic equation split into two branches, and the distribution function had a logarithmic infinity in each branch. He then investigated the general values of τ less than $\frac{1}{2}$; the break was made at this value because greater values of

τ correspond to next neighbour forces exceeding neighbour ones, an unlikely event in practice. This led to the following results: The largest frequency ν_L was $\{(4\alpha+8\gamma)/4\pi^2M\}^{1/2}$. The distribution function $g(\nu)$ had a value of infinity at $\nu=\nu_L\tau^{1/2}$, where τ was defined as before. There was a second infinity value at $\nu=\nu_L(1-\tau)^{-1/2}$ when $\tau<\frac{1}{8}$, and at $\nu=\nu_L(1+3\tau)/4\tau^{1/2}$ when $\tau>\frac{1}{8}$. As ν/ν_L tended to zero, $g(\nu)$ was linear in ν.

We may compare the results of applying (*a*) the method of calculation of sample modes and (*b*) Montroll's treatment as follows: (*a*) gives a knowledge of local values of the distribution function, but to get a good idea of the shape of the curve one must calculate for an immense number of points; (*b*) gives immediately a general idea of the trend of the curve, but local values are unreliable (except in special cases) unless one considers a large number of orders, the number necessary depending upon the convergence.

With the intention of obtaining a solution less laboriously Houston devised the treatment discussed in the next section.

5.3.3. *Houston's method of solution.*

Houston (1948) made calculations along special directions in the reciprocal lattice, e.g. in the simple cubic lattice along [001], [011] and [111], and then interpolated between these directions. He argued in this way: The secular equation for the vibration frequencies of a crystal lattice can be reduced to an equation of order $3N$ in the squares of the frequencies, where N is the number of different atoms per unit lattice cell. If N is 1, this gives an equation of order 3, and the solution gives the vibration frequencies in terms of the propagation vectors of the standing waves. These vectors fill the space in a Brillouin zone of the reciprocal lattice. The cubic equation can be solved numerically for a particular point, giving the frequency as a function of the coordinates of that point. By repeating this process for all points one could obtain a complete solution, but the work would be impossibly great. However, along a radius vector from the centre of the reciprocal lattice the frequency depends only on the distance. So by solving the cubic equation along this line one may say at what point the corresponding frequency has any particular value. The number of points lying close to this line, per unit solid angle, having values between ν and $\nu+d\nu$, will be given by $r^2(dr/d\nu)\,d\nu$, so that the distribution of frequencies within a narrow cone lying along this direction may be calculated. By solving along a number of such directions and summing, the distribution function may be obtained, and the closer the cones are together the better the result will be. Houston reduced the labour by making use of the symmetry of the problem: the frequency is a function of distance along the radius, and this function must be invariant to the operations of the symmetry group to which the crystal belongs. In the cubic lattice this divides the work by 48.

He investigated in this way the simple monatomic cubic lattice treated by Blackman (1937 a). He used two force constants, α between adjacent atoms when in motion in their line, and γ when the motion was perpendicular to this line, and, following Blackman, he took α as equal to 20γ. He examined three special directions as shown in fig. 2. He named the cones along the z axis, along $y=z$, and along $x=y=z$, the F_1, F_2 and F_3 cones respectively, and calculated the three functions $F=r^2\,dr/dq$ where q^2 was written for $m\omega^2/\alpha$. From F_1, F_2 and F_3 in the three chosen directions one had to deduce F for all directions. Since the crystal was cubic, Houston saw that he needed the three cubic harmonics of

lowest order of van der Lage and Bethe to set up the total distribution function. As a final result he found four very sharp peaks, the ordinates approaching infinity although the area under the curve was finite. Houston himself doubted the reality of all the peaks: he thought that the two derived from the F_1 cone were real, but had less confidence that the F_2 peaks were sharp, and thought that the F_3 peaks might well be spurious. He stated that his method tended to over-emphasize the peaks, and introduced ones probably spurious, but that using additional terms in the expansion might rectify these errors. The simple cubic lattice was only slowly convergent, and the body-centred or face-centred lattice would be more rapidly convergent and so yield a more satisfactory solution.

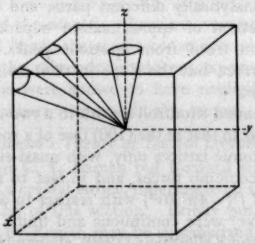


Fig. 2. Elementary cones in three typical directions.

5.3.4. *Criticism of the Montroll and Houston methods.*

Two papers have appeared which appraise these two treatments, and another paper carrying the work of Montroll a stage further.

The first was by Bowers and Rosenstock (1950). They chose to investigate a case which has little direct physical interest, viz. that of the vibrations normal to the plane, of a square two-dimensional simple lattice. It has the advantage, however, of having a definite analytical solution, so that it provides a criterion by which to judge other methods of solution. With α and γ as force constants for nearest and next nearest neighbours respectively, they plotted frequency distributions as ordinates against the square of frequency (relative to the square of the largest frequency) as abscissae for values of $\gamma/\alpha = 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$. In each case they found a frequency at which the density becomes infinite. For values of γ/α exceeding $\frac{1}{2}$ there was a frequency at which the density was discontinuous and doubled in value (e.g. at $(\nu/\nu_L)^2 = 0{\cdot}83$ for $\gamma/\alpha = \frac{3}{4}$). For small values of γ/α there was no qualitative difference from the case of $\gamma/\alpha = 0$, but only a small shift of the peak towards higher frequencies.

They then made the calculations again, first using Montroll's method and then Houston's.

The Montroll approximation was made up to the fifth order: it was recognized that this was insufficient to ensure faithful agreement, since they were trying to fit a curve having an infinite peak with a low order polynomial. The work necessary to produce a good fit would have been very laborious since the convergence was slow. The results were applied to the specific heat curve as a check, and above $T/\theta = 1$ agreement was obtained to the fourth significant figure.

The Houston approximation gave good agreement with the frequency distribution curve first calculated, up to the intensity maximum, but thereafter

there were wide discrepancies, culminating in a spurious infinite peak at the high-frequency limit of the spectrum. (This occurrence of spurious peaks had been recognized by Houston himself (1948).) The specific heat curve was matched fairly well, though less well than by Montroll's approximation. The Houston approximation was, however, less laborious to apply.

Nakamura (1950) independently made a similar study to that just described. He found, like Bowers and Rosenstock, that Houston's method gave results similar to Montroll's up to the first peak, but thereafter introduced spurious peaks. The less laborious nature of the Houston method appealed to him and, in his words, he "sought to salvage the method". The frequency distribution function consisted of analytically different parts, and by separating these and applying Houston's method of approximation separately to them he found a frequency distribution freed from spurious peaks. The positions of the singular peaks were correct, but the slope in their neighbourhood was steeper than it should have been.

Smollett (1952) extended Montroll's work to a two-dimensional ionic lattice, having a structure similar to that of the (100) face of a rocksalt crystal. Montroll had worked with monatomic lattices only, with quasi-elastic forces. Smollett's calculations applied to coulomb forces, and in fact to any forces, provided the first three derivatives of $f^2(=4\pi^2 M\nu^2)$ with respect to ϕ_1 and ϕ_2 (the two phase components for the wave) were continuous and finite. He used the methods of Blackman (1935 a, b) and Kellermann (1940), allowing him to calculate the exact density of vibration frequencies at special points, but not to give an analytic expression for the whole spectrum.

Considering the phase-maps obtained (contours for equal frequencies on a (ϕ_1, ϕ_2) plot) on the low-frequency map at the points $(0, \pi)$ and $(\pi, 0)$, an infinite density of logarithmic type was found. At similar points in the high-frequency map the density was finite, but a logarithmic infinite density was found somewhat short of the high-frequency limit. These two infinite peaks lie at about the same values of ν/ν_L as found by Montroll (1947) for $\tau = \frac{1}{3}$. Smollett found in addition two finite peaks, one in each of the high-frequency and low-frequency branches. The former was at the limiting frequency and the latter at about four-fifths of it. Neither of these subsidiary peaks occurs with quasi-elastic forces.

The spectrum of the three-dimensional ionic crystal is likely to have considerable similarity to the case investigated by Smollett, and there are likely to be infinite peaks of similar character, accounting for peaks of continuum of line-like character, in the Raman spectrum of second order.

5.4. *Anharmonicity of Vibrations*

In two papers (1951, 1952), K. S. Krishnan and Roy have studied the anharmonicity of the vibrations of alkali halide crystals, both face-centred and body-centred. Using Born's model, they derive an expression for the anharmonicity by considering the coefficient of the fourth power of the displacement.

In the first paper they deal with the homogeneous vibration ν_0 due to the oscillations of the lattices of the alkali and the halide ions with respect to one another. The anharmonicity of this vibration was found to depend upon its direction, being largest along (111). Incidentally, they showed that there is a

large contribution to the forces affecting this vibration due to the polarization arising from the small relative displacements of the positive and negative lattices. For example, in rocksalt the force arising in this way is nearly half that due to the interactions with other ions, and of opposite sign, so that the polarization force is of the same order as the resultant force.

In the second paper, general expressions for the relative amplitudes of the positive and negative ions were calculated for any normal vibration of alkali halide crystals, together with the frequencies. A detailed study was made of the low-frequency acoustic vibrations. For these also it was found that the polarization field was important, affecting the relative amplitudes but not the frequencies.

From the frequencies of the acoustic vibrations were calculated the velocities of the acoustic waves, and hence the elastic constants, which were found to be in general agreement with the values deduced from experiment.

The acoustic vibrations were shown to have negligible anharmonicity, so differing from the homogeneous vibration in this respect.

5.5. *Raman's Theory of Lattice Dynamics*

In a series of papers Sir C. V. Raman put forward an atomistic view of the nature of the vibrations of a crystal lattice, differing from that of Born and others described in the earlier part of this section. Briefly put, the ideas underlying the theory as put forward by Raman may be summarized as follows:

In the interaction of light with the atoms composing a crystal one has to do with an incoherent scattering in which quanta of energy are transferred from the radiation field to individual elements of volume of molecular dimensions in the crystal. In ionic or molecular crystals lines appear in the Raman spectrum having nearly the same frequencies as with the same substance when fluid. Overtones and summations of fundamental vibration frequencies appear in the Raman spectrum of second order, showing that the amplitudes of vibration are not negligible compared with the distances between the atoms, and optical anharmonicity is at work. This shows that the normal vibrations of the atoms are concerned, and the problem is similar to that of the vibrations of polyatomic molecules. The boundaries of the crystal are then irrelevant. Attention is to be directed to the atoms vibrating with the same frequency and with the same or opposite phases. Rayleigh (1926) stated that "all particles in a normal vibration have at any instant the same or opposite phase of vibration".

These ideas led Raman to postulate that any mode of vibration should either be invariant to a unit lattice translation or the amplitude should remain unaltered but the phase be reversed. This yields eight different cases. If there are p atoms in a unit cell there are $3p$ equations for the modes and frequencies of vibration for each case, and $3p$ independent coordinates. This means that there are $24p-3$ normal modes, since there are 3 translations. $3p-3$ modes are invariant to the lattice translations, while the $21p$ remaining have alternating phases along at least one of the axes. This comes approximately to the same problem as that of the vibrations of a 'supercell' containing $8p$ atoms, i.e. one having twice the unit translation in each direction. The opposite phases implicit in the $21p$ modes would lead, in first approximation, to no variation in the polarizability, so that they would be inactive in the Raman spectrum of first order, and only the $3p-3$ modes could give rise to fundamental vibrations. This would not be the case for higher order Raman spectra.

Considering an actual crystal which consists of N supercells, since three degrees of freedom have been allowed for each supercell there are $3N$ degrees of freedom unallocated, and these are attributed to the elastic vibrations of the crystal, with limiting wavelengths in the crystal equal to the edges of the supercell. Raman therefore expects a high frequency (I.R. and R.E.) discrete spectrum due to the interatomic vibrations, and an elastic continuum spectrum of a lower frequency range.

Sir C. V. Raman applied these ideas to elucidate the R.E. and I.R. spectra of a number of crystals, notably diamond, several alkali halides, and the very detailed infra-red observations of Barnes, Brattain and Seitz (1935) on magnesium oxide. Many experimental papers have appeared in support of Raman's views by his co-workers, notably R. S. Krishnan. These are referred to in the experimental section of this report.

Raman has vigorously defended his views and has sharply criticized the assumption of the cyclic lattice. The matter goes far outside the subject of the present report, since all properties of crystals depending on the motions of the atoms are affected, from specific heats to 'extra spots' in Laue photographs. For the latest formulation of Raman's views the interested reader should consult Raman (1951); for a rigorous treatment of the diatomic finite chain and the comparison with the results deduced from the cyclic lattice see Born (1942).

§6. ALKALI HALIDES: EXPERIMENTAL

The early experiments failed to show any R.E. with these crystals, because the homogeneous vibration is active only in the I.R. Rasetti succeeded in obtaining a Raman spectrum, using the elegant technique devised by him (Fermi and Rasetti 1931) employing the mercury resonance line as exciter, referred to earlier. This technique allowed him to employ a very long exposure, without fogging of the plate or great halation from the exciting line. The spectrum consisted of a band of continuum with peaks superposed, one being so sharp as to have almost a line-like character. All subsequent experiments on such second order Raman spectra of the alkali halides have used Rasetti's technique.

Later a series of papers has appeared due to R. S. Krishnan, e.g. NaCl (1947), KBr (1948). In 1943 R. S. Krishnan sought to explain Rasetti's results in terms of Raman's theory, finding in Rasetti's spectrum the nine discrete lines expected. In 1947 he published the results of his own experiments, finding 24 Stokes lines, his criterion for their reality being the coincidences of peaks in two microphotometer records. Of the 24 lines, 19 were attributed to the second order spectrum, and the remaining five to a third order spectrum (normally forbidden by the symmetry of the crystal). The sharpest peak at 235 cm^{-1} was regarded as a line, and not a peak in a continuum.

A special investigation of this point was made by Welsh, Crawford and Staple (1949). They photographed the Raman spectrum of rocksalt with a dispersion about three times that used by Krishnan, and checked the accuracy of the intensity measurements and the effects of parasitic radiation by photographing also the rotational Raman spectrum of nitrogen under precisely similar conditions. They concluded that the spectrum consisted of a continuum with superimposed maxima of relatively small intensity. The measured half-intensity width of the peak at 234 cm^{-1} was 8 cm^{-1}, compared with 1 cm^{-1} for a neighbouring mercury line. Their observations convinced them of the practical

impossibility of interpreting the spectrum in terms of nine discrete lines, but they considered that their spectrum was, at least qualitatively, a striking confirmation of the Born theory.

Menzies and Skinner (1948a) published an account of *polarized* Raman spectra of second order, obtained for NaCl, KBr and KCl. The first two substances had spectra which were polarized, except for the strong peaks at 232 cm^{-1} and 127 cm^{-1} respectively, which were strongly depolarized. KCl on the other hand gave a spectrum which was everywhere polarized and very much weaker than those for the first two substances named. They based their explanation of this difference on the effect that the mass-ratios of the respective cation and anion had on the shape of the frequency–phase diagram (see §5.2), thus being guided by the Born theory.

The same authors (1948b) proceeded to test this by examining the polarized spectra of two other substances, one having a large mass-ratio of the ions (NaBr) and one having, like KCl, a mass-ratio of the ions close to unity (RbBr). They found that NaBr behaved like NaCl and KBr, while RbBr behaved like KCl, as expected.

The Russian authors incline to the continuum origin of the alkali halide spectra.

Much still remains to be examined: only a small number of alkali halides have been examined experimentally, and so far none of the body-centred type. The Raman spectrum of MgO would be interesting. The infra-red data are also very incomplete. Burstein, Oberly and Plyler (1948), and independently Willmott, failed to confirm the detailed infra-red absorption spectrum of MgO found by Barnes, Brattain and Seitz. Born's continuum theory as first worked out did not account for the sharpness of some of the peaks, but the logarithmic infinities of Montroll (1943) may now supply this deficiency. Apart from these there is also the possibility of sharp peaks of intensity not necessarily corresponding to logarithmic infinities, as shown by Smollett (1952).

§7. CONCLUDING REMARKS

It has not been possible to touch upon many interesting problems upon which much work has been done in recent years, such as the Raman spectrum of diamond, the effect of temperature changes, especially change of state and the effect of passing through a transition point.

It is hoped that the bibliography which follows may to some extent make this deficiency good. In it the references are given, followed by very brief remarks intended to indicate the interest of the respective papers from the point of view of this report.

Note added in proof (2.3.53). Through the courtesy of M. Blackman I have seen a manuscript by L. van Hove, of the Institute for Advanced Study, Princeton, which is important in connection with the subject matter of §5.3. Van Hove shows that, assuming harmonicity for the interatomic forces, as a consequence of the periodic structure the frequency distribution function has analytic singularities. In the general case the nature of these singularities depends only on the number of dimensions of the crystal.

For a two-dimensional 'crystal' the distribution function has logarithmically infinite peaks. This is not so for a three-dimensional crystal, so that the expectation expressed by Smollett (1952) is not fulfilled. The argument is general: it could only be upset by assumption of special forces between particles. However, the sharp peaks calculated, for example, by Iona (1941) for KCl, and by Leighton (1948) for face-centred cubic crystals, are not placed in doubt by van Hove's argument.

§8. REFERENCES

8.1. *Theory*

AGGARWAL, R. R., and SAKSENA, B. D., 1951, *J. Chem. Phys.*, **19**, 1480.
Lattice frequencies in molecular mix-crystals.

BARNES, R. B., BRATTAIN, R. R., and SEITZ, F., 1935, *Phys. Rev.*, **48**, 582.
MgO.

BLACKMAN, M., 1933, *Z. Phys.*, **86**, 421.
Subsidiary I.R. maxima calculated with 1- and 3-dimensional lattices.

BLACKMAN, M., 1934, *Phil. Mag.*, **18**, 425.
Coupling of internal with external vibrations.

BLACKMAN, M., 1935 a, *Proc. Roy. Soc.* A, **148**, 365.
Frequency distribution and specific heats.

BLACKMAN, M., 1935 b, *Proc. Roy. Soc.* A, **149**, 117.
Continuation of preceding paper.

BLACKMAN, M., 1936, *Phil. Trans. Roy. Soc.* A, **236**, 103.
2- and 3-dimensional diatomic crystals: NaCl and KCl.

BLACKMAN, M., 1937 a, *Proc. Roy. Soc.* A, **159**, 416.
Cubic lattice with special forces.

BLACKMAN, M., 1937 b, *Proc. Camb. Phil. Soc.*, **33**, 94.
Low frequencies in diatomic cubic lattice.

BLACKMAN, M., 1938, *Rep. Prog. Phys.*, **8**, 11 (London: Physical Society).
Frequency distributions and specific heats.

BLACKMAN, M., 1952, *Proc. Phys. Soc.* A, **65**, 394.
Non-cubic ionic lattice-frequencies depend on wave direction.

BORN, M., 1942, *Proc. Phys. Soc.*, **54**, 362.
Diatomic linear lattice.

BORN, M., 1945, *Rev. Mod. Phys.*, **7**, 245.
Lattice dynamics and pyroelectricity.

BORN, M., and BLACKMAN, M., 1933, *Z. Phys.*, **82**, 551.
Secondary I.R. maxima calculated with 1-dimensional lattice.

BORN, M., and BRADBURN, M., 1947, *Proc. Roy. Soc.* A, **188**, 161.
Rocksalt.

BORN, M., and VON KÁRMÁN, T., 1912, *Phys. Z.*, **13**, 297.
Equations for cubic lattice.

BOWERS, W. A., and ROSENSTOCK, H. B., 1950, *J. Chem. Phys.*, **18**, 1056.
Comparison of Montroll and Houston approximations with rigorous case.

BURSTEIN, E., OBERLY, J. J., and PLYLER, E. K., 1948, *Proc. Ind. Acad. Sci.* A, **28**, 388.
Effect of electrical anharmonicity.

CABANNES, J., 1932, *Ann. Phys., Paris*, **18**, 285.
Symmetry of molecules and R.E. selection rules.

COUTURE, J., 1947, *La Rev. Sci.*, **85**, 86.
Frequencies in ionic crystals.

COUTURE, L., and MATHIEU, J. P., 1948, *C. R. Acad. Sci., Paris*, **226**, 1596.
Circular polarization.

DEHLINGER, H., 1914, *Phys. Z.*, **15**, 276.
Cubic lattice equations.

FINE, P. C., 1939, *Phys. Rev.*, **56**, 335.
Body-centred lattice numerical calculations: two steep maxima on a continuum.

GERSTENKORN, H., 1952, *Ann. Phys., Lpz.*, **10**, 80.
Elastic waves in cubic lattices.

HALFORD, J. O., 1951, *J. Chem. Phys.*, **19**, 1375.
1- and 3-dimensional cases.

HALFORD, J. O., 1952, *J. Chem. Phys.*, **20**, 822.
Simple cubic lattice: no infinities.

HORNIG, D. F., 1948, *J. Chem. Phys.*, **16**, 1063.
Internal and external vibrations and their interactions: effects of symmetry.

HOUSTON, W. V., 1948, *Rev. Mod. Phys.*, **20**, 161.
Averaging of directional calculations.
HUANG, K., 1951, *Proc. Roy. Soc.* A, **208**, 352.
Interaction between radiation field and ionic crystals.
IONA, M., 1941, *Phys. Rev.*, **60**, 822.
Frequency distribution in KCl.
KELLERMANN, E. W., 1940, *Phil. Trans. Roy. Soc.*, **238**, 513.
Frequency distribution in NaCl.
KRISHNAN, Sir K. S., and ROY, S. K., 1951, *Proc. Roy. Soc.* A, **207**, 447.
Anharmonicity of alkali halide fundamental vibrations.
KRISHNAN, Sir K. S., and ROY, S. K., 1952, *Proc. Roy. Soc.* A, **210**, 481.
Anharmonicity of alkali halide low-frequency vibrations.
LEIGHTON, R. B., 1948, *Rev. Mod. Phys.*, **20**, 165.
Frequency distribution calculation aided by models.
LOINGER, A., 1950, *Nuovo Cim.*, **7**, 939.
Intensities of internal and external vibration lines.
LONSDALE, K., 1942, *Nature, Lond.*, **149**, 698.
Crystal dynamics of rocksalt.
LÖWDIN, P. O., 1950, *J. Chem. Phys.*, **18**, 365.
Binding energies between alkali and halogen ions.
LYDDANE, R. H., and HERZFELD, K. F., 1938, *Phys. Rev.*, **54**, 846.
Ionic cubic lattice : numerical calculations.
MATHIEU, J. P., 1945, *Spectre de vibration et symétrie des molécules et des cristaux* (Paris : Hermann).
MONTET, G. L., KELLER, S. P., and MAYER, J. E., 1952, *J. Chem. Phys.*, **20**, 1057.
Binding energies between alkali and halogen ions (density matrix method).
MONTROLL, E. W., 1942, *J. Chem. Phys.*, **10**, 218.
Moment-matrix method of solution : application to 2-dimensional lattice.
MONTROLL, E. W., 1943, *J. Chem. Phys.*, **11**, 481.
Simple cubic lattice : sharp maxima.
MONTROLL, E. W., 1947, *J. Chem. Phys.*, **15**, 575.
Square lattice : two infinities.
MONTROLL, E. W., and PEASLEE, D. C., 1944, *J. Chem. Phys.*, **12**, 98.
Body-centred cubic lattice.
MOTULEVICH, G. P., 1947, *Izv. Akad. Nauk, S.S.S.R.*, Ser. Fiz. II, 390.
Frequencies in terms of elastic constants.
MOTULEVICH, G. P., and TUROVTSEVA, Z. M., 1950, *J. Exp. Theor. Phys., U.S.S.R.*, **20**, 234.
Calculations of calcite frequencies, intensities and depolarizations.
NAKAMURA, T., 1950, *Prog. Theor. Phys. (Kyoto)*, **5**, 213.
Comparison of Montroll's and Houston's methods of calculation and extension of latter.
ODELEVSKIĬ, V. I., 1950, *Izv. Akad. Nauk, S.S.S.R.*, Ser. Fiz. 14, 232.
Polyhedron of anions around cation, or vice versa, taken as the elementary vibrator in a cubic ionic crystal.
NIERENBERG, W. A., 1951, *J. Chem. Phys.*, **19**, 659.
Frequency distribution : matrix method.
PLACZEK, G., 1934, *Handbuch der Radiologie*, **6**, pt. 2, 205.
Monograph on theory.
PORFIRÉVA, N. N., 1949, *J. Exp. Theor. Phys., U.S.S.R.*, **19**, 692.
2- and 3-dimensional molecular crystals.
PORFIRÉVA, N. N., 1950, *J. Exp. Theor. Phys., U.S.S.R.*, **20**, 97.
Frequencies and selection-rules for molecular crystals.
POULET, H., 1952, *C.R. Acad. Sci., Paris*, **234**, 2185.
Dependence of frequency on angle between directions of vibration and of propagation.
RAMAN, Sir C. V., 1928, *Ind. J. Phys.*, **2**, 387.
Announcement of discovery of the effect.
RAMAN, Sir C. V., 1941 a, *Proc. Ind. Acad. Sci.* A, **13**, 1.
Optical vibrations excited by photons.

RAMAN, Sir C. V., 1941 b, *Proc. Ind. Acad. Sci.* A, **14,** 459.
Vibrations of lattice and superlattice cells, and specific heats.
RAMAN, Sir C. V., 1943, *Proc. Ind. Acad. Sci.* A, **18,** 237.
Lattice dynamics : phase either same or alternate in successive cells.
RAMAN, Sir C. V., 1947, *Proc. Ind. Acad. Sci.* A, **26,** 339, 356, 370, 383, 391, 396.
Theory of lattice dynamics with applications to diamond, NaCl, MgO, LiF, NaF and KCl.
RAMAN, Sir C. V., 1948, *C.N.R.S. Colloquium, Bordeaux,* 1.
Lattice dynamics : general review of theories.
RAMAN, Sir C. V., 1951, *Proc. Ind. Acad. Sci.* A, **34,** 61, 141.
(1) Monochromatic nature of the vibrations in crystals, with consideration of overtones and summations. (2) Specific heats.
RAŠKIN, S. S., SECHKAREV, A. V., and SKRIPOV, F., 1949, *Dokl. Akad. Nauk., S.S.S.R.,* **66,** 837.
Molecular lattices.
RAYLEIGH, Lord, 1926, *Theory of Sound,* Vol. I, p. 107 (London : Macmillan).
ROBERTS, S., 1950, *Phys. Rev.,* **77,** 258.
Alkali halide frequencies and dielectric polarization.
SAKSENA, B. D., 1940, *Proc. Ind. Acad. Sci.* A, **11,** 229.
Depolarizations according to classes.
SAYRE, E. V., and BEAVER, J. J., 1950, *J. Chem. Phys.,* **18,** 584.
Numerical calculation of frequencies of crystals of NaH and NaD using punched cards.
SKRIPOV, F., 1949, *Dokl. Akad. Nauk, S.S.S.R.,* **66,** 1075.
Effect of symmetry on limiting frequencies of molecular lattices.
SMITH, H. M. J., 1948, *Phil. Trans. Roy. Soc.* A, **241,** 105.
Numerical calculation of frequency distribution and R.E. of diamond.
SMOLLETT, M., 1952, *Proc. Phys. Soc.* A, **65,** 109.
Frequency-spectrum of 2-dimensional ionic lattice.
SZIGETI, B., 1950, *Proc. Roy. Soc.* A, **204,** 51.
Compressibility and absorption frequencies of ionic crystals.
THEIMER, O., 1948, *Proc. Ind. Acad. Sci.* A, **28,** 505.
Influence of cation on intensity of R.E. in alkali halides.
THEIMER, O., 1951, *Proc. Phys. Soc.* A, **64,** 1012.
Relative intensity of the fundamental Raman frequency and of Brillouin components in diamond.
THEIMER, O., 1952, *Proc. Phys. Soc.* A, **65,** 38.
Photoelastic properties and R.E. in crystals.
TISZA, L., 1933, *Z. Phys.,* **82,** 48.
Group theoretical treatment of vibrations.
TOLPYGO, K. B., 1950, *J. Exp. Theor. Phys., U.S.S.R.* **20,** 497.
Alkali halide lattice of deformed ions.
WALNUT, T. H., 1952, *J. Chem. Phys.,* **20,** 58.
Selection rules for vibrations of complex crystals.

8.2. *Experimental Technique*

ANANTHAKRISHNAN, R., 1937, *Proc. Ind. Acad. Sci.,* **5,** 76.
Complementary filters.
BÄR, R., 1929, *Nature, Lond.,* **124,** 692.
Powder technique.
GERLACH, W., 1930, *Ann. Phys., Lpz.,* **5,** 196.
Powder technique.
GÜBER, W., and RIGGERT, K. H., 1951, *Z. Naturforschung,* **8,** 464.
Complementary interference filters.
HARRAND, M., 1948, *J. Phys. Radium,* **9,** 81.
Monochromator powder technique.
KORSHUNOV, A. V., and SELKIN, V. A., 1950, *J. Tech. Phys., U.S.S.R.,* **20,** 745.
Strong illumination.

KRISHNAMURTI, P., 1930, *Ind. J. Phys.*, **5**, 1.
Powder technique.
LANDSBERG, G., and MANDELSTAM, L., 1928, *Naturwissenschaften*, **16**, 557.
Early experiments on quartz and calcite.
MAYENCE, J., 1951, *J. Phys. Radium*, **12**, 744.
Low temperatures.
MÉDARD, L., 1934, *C. R. Acad. Sci., Paris*, **199**, 421.
Complementary filters.
MENZIES, A. C., 1929, *Nature, Lond.*, **124**, 511.
Early powder technique.
MENZIES, A. C., and MILLS, H. R., 1935, *Proc. Roy. Soc.* A, **148**, 407.
Low temperature apparatus.
RAMAN, Sir C. V., and KRISHNAN, Sir K. S., 1928, *Nature, Lond.*, **121**, 501.
Discovery of the R.E.
RASETTI, F., 1932, *Nuovo Cim.*, **9**, 72.
Mercury resonance-line technique.
REITZ, A. W., 1940, *Z. Phys. Chem.* B, **46**, 181.
Monochromator powder technique.
TAYLOR, W. J., SMITH, A. L., and JOHNSTON, H. L., 1951, *J. Opt. Soc. Amer.*, **41**, 91.
Low temperature apparatus to 20°K.
VODAR, B., JARDILLIER, Y., and MAYENCE, J., 1946, *C. R. Acad. Sci., Paris*, **222**, 1493.
Low temperature apparatus.

8.3. *Inorganic (other than Alkali Halides)*

AYNARD, R., 1952, *C. R. Acad. Sci., Paris*, **234**, 2352.
Calcite : interpretation of anomalies.
CHAPELA, J., 1948, *C. R. Acad. Sci., Paris*, **226**, 1814.
Single crystal, $NH_4H_2PO_4$.
CHAPELLE, J., 1948, *C. R. Acad. Sci., Paris*, **226**, 401.
Water of crystallization in Seignette salt.
CHAPELLE, J., 1949, *C. R. Acad. Sci., Paris*, **228**, 755.
Low frequencies in ammonium and K phosphates.
CHAPELLE, J., and GALY, A., 1952, *J. Phys. Radium*, **13**, 111.
Seignette salt.
COUTURE, L., 1947, *Ann. Phys., Paris*, **2**, 5.
Intensities and depolarizations of lines and structures of seven ionic crystals, either carbonates, nitrates or sulphates.
COUTURE, L., 1948, *J. Phys. Radium*, **9**, 84.
Effect of crystal field on complex ion : anhydrite.
COUTURE, L., and MATHIEU, J. P., 1947, *C. R. Acad. Sci., Paris*, **225**, 1140.
K_2SnCl_6, nitrates of Sr, Ba, Pb; $NaClO_3$, $NaBrO_3$.
COUTURE, L., and MATHIEU, J. P., 1948 a, *C. R. Acad. Sci., Paris*, **226**, 1261.
Ammonium chloride.
COUTURE, L., and MATHIEU, J. P., 1948 b, *Ann. Phys., Paris*, **3**, 521.
Cubic crystals.
COUTURE, L., and MATHIEU, J. P., 1948 c, *Proc. Ind. Acad. Sci.* A, **28**, 401.
Ammonium chloride and ammonium bromide : symmetry of ammonium ion.
COUTURE, L., and MATHIEU, J. P., 1949, *C. R. Acad. Sci., Paris*, **228**, 555.
Acid fluoride of K and structure of ion FHF.
COUTURE, L., and RICHE, N., 1949, *J. Phys. Radium*, **10**, 151.
Acid sulphates of K and Am. Light on structures previously unknown.
COUTURE-MATHIEU, L., 1949, *C. R. Acad. Sci., Paris*, **229**, 1215.
Potassium bicarbonate : CO_3 ions linked by H bonds.
COUTURE-MATHIEU, L., 1950, *J. Phys. Radium*, **11**, 541.
$KHCO_3$: structure previously unknown.
COUTURE-MATHIEU, L., and MATHIEU, J. P., 1950, *C.R. Acad. Sci., Paris*, **231**, 839.
Variation in frequency with direction of waves : iodic acid.
COUTURE-MATHIEU, L., and MATHIEU, J. P., 1951, *J. Phys. Radium*, **12**, 826.
Fluosilicate ions.

COUTURE-MATHIEU, L., POULET, H., and MATHIEU, J. P., 1952, *C. R. Acad. Sci., Paris*, **234**, 1761.
Abnormal depolarizations in $NaClO_3$, $NaBrO_3$.

GALY, A., 1951 a, *J. Phys. Radium*, **12**, 827.
$NaH_2PO_4 . 2H_2O$.

GALY, A., 1951 b, *C.R. Acad. Sci., Paris*, **233**, 1181.
$MnCl_2 . 4H_2O$.

GIULOTTO, L., and OLIVELLI, G., 1948, *Nuovo Cim.*, **5**, 52.
Polarization of low frequencies in calcite.

GROSS, E. F., and VALKOV, V. I., 1950, *Dokl. Akad. Nauk, S.S.S.R.*, **74**, 453.
Ice.

GROSS, E. F., and VALKOV, V. I., 1951, *Dokl. Akad. Nauk, S.S.S.R.*, **81**, 761.
Heavy ice.

HIEBERT, G. L., and HORNIG, D. F., 1952, *J. Chem. Phys.*, **20**, 918.
Solution of HCl in DCl.

KASTLER, A., 1948 a, *Proc. Ind. Acad. Sci.* A, **28**, 349.
External vibrations in calcite : coupling between translations and rotations.

KASTLER, A., 1948 b, *C. R. Acad. Sci., Paris*, **227**, 1024.
Coupling of external vibrations.

KRISHNAN, R. S., 1945 a, *Proc. Ind. Acad. Sci.* A, **22**, 182.
Calcite (2nd order).

KRISHNAN, R. S., 1945 b, *Proc. Ind. Acad. Sci.* A, **22**, 274.
Gypsum (2nd order).

KRISHNAN, R. S., 1945 c, *Proc. Ind. Acad. Sci.* A, **22**, 329.
Quartz (2nd order).

KRISHNAN, R. S., 1947 a, *Proc. Ind. Acad. Sci.* A, **26**, 399.
Diamond.

KRISHNAN, R. S., 1947 b, *Proc. Ind. Acad. Sci.* A, **26**, 432.
Ammonium chloride : temperature.

KRISHNAN, R. S., 1947 c, *Proc. Ind. Acad. Sci.* A, **26**, 450.
Crystalline alumina.

KRISHNAN, R. S., 1947 d, *Proc. Ind. Acad. Sci.* A, **26**, 460.
Topaz.

KRISHNAN, R. S., 1948 a, *Proc. Ind. Acad. Sci.* A, **27**, 321.
Ammonium bromide.

KRISHNAN, R. S., 1948 b, *Proc. Ind. Acad. Sci.* A, **28**, 307.
Diamond.

KRISHNAN, R. S., 1950 a, *Proc. Ind. Acad. Sci.* A, **31**, 435.
Aragonite.

KRISHNAN, R. S., 1950 b, *Proc. Ind. Acad. Sci.* A, **32**, 105.
Anhydrite : different directions.

KUMARI, C. S., 1948, *Proc. Ind. Acad. Sci.* A, **28**, 500.
Sodium chlorate.

KUMARI, C. S., 1950 a, *Curr. Sci.*, **19**, 204.
Barium chlorate.

KUMARI, C. S., 1950 b, *Proc. Ind. Acad. Sci.* A, **32**, 177.
Potassium chlorate : temperature.

MARIÉE, M., and MATHIEU, J. P., 1946, *C. R. Acad. Sci., Paris*, **223**, 147.
Effect of compression on quartz.

MATHIEU, J. P., 1948, *J. Phys. Radium*, **9**, 83.
Anisotropy in cubic crystals.

MATHIEU, J. P., 1949, *C. R. Acad. Sci., Paris*, **229**, 1068.
Hydrated nitrates.

MATHIEU, J. P., 1952, *C. R. Acad. Sci., Paris*, **234**, 2272.
$NaClO_3$.

MATHIEU, J. P., AGUIRRE, R. M., and COUTURE-MATHIEU, L., 1951, *C. R. Acad. Sci., Paris*, **232**, 318.
Ammonium chloride : low temperature.

MATHIEU, J. P., and COUTURE-MATHIEU, L., 1950, *C. R. Acad. Sci., Paris*, **230**, 1054.
Potassium hydrofluoride.

MATHIEU, J. P., and COUTURE-MATHIEU, L., 1951 a, *C. R. Acad. Sci., Paris*, **233**, 32.
Zinc blende.
MATHIEU, J. P., and COUTURE-MATHIEU, L., 1951 b, *C. R. Acad. Sci., Paris*, **233**, 1595.
Orientation of ammonium group in ammonium halides.
MATHIEU, J. P., and COUTURE-MATHIEU, L., 1952 a, *J. Phys. Radium*, **13**, 271.
$LiClO_4 . 3H_2O$: three frequencies sensitive to wave directions.
MATHIEU, J. P., and COUTURE-MATHIEU, L., 1952 b, *C. R. Acad. Sci., Paris*, **234**, 1961.
Quartz : frequencies, intensities and depolarizations of certain lines depend on separations of longitudinal and elastic waves.
MATOSSI, F., 1951, *J. Chem. Phys.*, **19**, 1543.
Rutile.
NARAYANAN, P. S., 1948 a, *Proc. Ind. Acad. Sci.* A, **28**, 463.
Stokes/anti-Stokes intensities.
NARAYANAN, P. S., 1948 b, *Proc. Ind. Acad. Sci.* A, **28**, 469.
Ammonium dihydrogen phosphate.
NARAYANAN, P. S., 1950 a, *Proc. Ind. Acad. Sci.* A, **32**, 279.
Rutile.
NARAYANAN, P. S., 1950 b, *Curr. Sci.*, **19**, 113.
Barytes.
NARAYANAN, P. S., 1951 a, *Proc. Ind. Acad. Sci.* A, **33**, 97.
Stokes/anti-Stokes intensities.
NARAYANAN, P. S., 1951 b, *Curr. Sci.*, **20**, 94.
Apophyllite.
NARAYANASWAMY, P. K., 1947 a, *Proc. Ind. Acad. Sci.* A, **26**, 511.
Calcite : external frequencies more affected by temperature than internal ones.
NARAYANASWAMY, P. K., 1947 b, *Proc. Ind. Acad. Sci.* A, **26**, 521.
Quartz : temperature effects.
NARAYANASWAMY, P. K., 1948 a, *Proc. Ind. Acad. Sci.* A, **27**, 311.
Ice.
NARAYANASWAMY, P. K., 1948 b, *Proc. Ind. Acad. Sci.* A, **28**, 40.
Barytes : temperature effect.
NARAYANASWAMY, P. K., 1948 c, *Proc. Ind. Acad. Sci.* A, **28**, 417.
β-quartz: α-β transition.
PRESS, D. C., 1950, *Proc. Ind. Acad. Sci.* A, **31**, 56.
Fluorspar : temperature effect.
PROD'HOMME, L., 1951, *C. R. Acad. Sci., Paris*, **233**, 303.
Glass.
RAMANATHAN, K. G., 1948, *Proc. Ind. Acad. Sci.* A, **28**, 459.
ZnS structures.
RANK, D. H., and DOUGLAS, A. E., 1948, *J. Opt. Soc. Amer.*, **38**, 966.
Optical glass.
SAKSENA, B. D., and NARAIN, H., 1949 a, *Nature, Lond.*, **164**, 583.
β-quartz.
SAKSENA, B. D., and NARAIN, H., 1949 b, *Proc. Ind. Acad. Sci.* A, **30**, 128.
β-quartz.
SANTANGELO, M., 1951, *Ricerca Sci.*, **21**, 1391.
Gypsum.
SMITH, A. L., KELLER, W. E., and JOHNSTON, H. L., 1951, *J. Chem. Phys.*, **19**, 189.
Solid and liquid condensed nitric oxide.
VASSAS-DUBUISSON, C., 1948, *J. Phys. Radium*, **9**, 91.
$NaNO_3$: temperature effect.
VASSAS-DUBUISSON, C., 1949, *C. R. Acad. Sci., Paris*, **228**, 1712.
Sodium thiosulphate.
VASSAS-DUBUISSON, C., 1951, *C. R. Acad. Sci., Paris*, **233**, 374.
$Li_2SO_4 . H_2O$.
WAGNER, E. L., and HORNIG, D. F., 1950 a, *J. Chem. Phys.*, **18**, 296.
Ammonium chloride and deutero-ammonium chloride.
WAGNER, E. L., and HORNIG, D. F., 1950 b, *J. Chem. Phys.*, **18**, 305.
Ammonium bromide and deutero-ammonium bromide.

8.4. *Alkali Halides: Experimental*

FERMI, E., and RASETTI, F., 1931, *Z. Phys.*, **71**, 689.
NaCl: Hg resonance technique.

GROSS, E. F., PAVINSKIĬ, P. P., and STEKHANOV, A. I., 1949, *Dokl. Akad. Nauk, S.S.S.R.*, **68**, 27.
NaCl.

GROSS, E. F., and STEKHANOV, A. I., 1947, *Nature, Lond.*, **160**, 568.
Temperature dependence in NaCl.

KRISHNAN, R. S., 1943, *Proc. Ind. Acad. Sci.* A, **18**, 298.
Interpretation of NaCl spectrum.

KRISHNAN, R. S., 1947, *Proc. Ind. Acad. Sci.* A, **26**, 419.
NaCl.

KRISHNAN, R. S., 1948, *Proc. Ind. Acad. Sci.* A, **28**, 296.
KBr.

KRISHNAN, R. S., and NARAYANAN, P. S., 1949, *Nature, Lond.*, **163**, 570.
KBr.

MENZIES, A. C., and SKINNER, J., 1948 a, *J. Phys. Radium*, **9**, 93.
Polarized 2nd order R.E. in NaCl, KBr, KCl.

MENZIES, A. C., and SKINNER, J., 1948 b, *C.N.R.S. International Colloquium, No. 7, Bordeaux*, **60**.
Polarized 2nd order R.E. in NaBr, RbBr.

STEKHANOV, A. I., 1950, *J. Exp. Theor. Phys., U.S.S.R.*, **20**, 330.
KCl.

STEKHANOV, A. I., and PETROVA, M. L., 1949, *J. Exp. Theor. Phys., U.S.S.R.*, **19**, 1108.
NaBr and KI.

WELSH, H. L., CRAWFORD, M. F., and STAPLE, W. J., 1949, *Nature, Lond.*, **164**, 737.
NaCl: continuum not line.

8.5. *Organic Solids*

BISHUI, B. M., 1948 a, *Ind. J. Phys.*, **22**, 319.
1, 2- and 1, 1-dichloroethanes in solid and liquid states.

BISHUI, B. M., 1948 b, *Ind. J. Phys.*, **22**, 334.
n-propyl bromide and ethylene chlorhydrin in solid and liquid states.

BISHUI, B. M., 1948 c, *Ind. J. Phys.*, **22**, 447.
Acetyl chloride, acetyl bromide and ethyl bromide at low temperatures.

BROWN, J. K., SHEPPARD, N., and SIMPSON, D. M., 1950, *Discussions of the Faraday Society*, No. 9, 261.
Normal paraffins.

DAASCH, L. W., and SMITH, D. C., 1951, *J. Chem. Phys.*, **19**, 22.
Trimethylphosphine oxide.

FRÜHLING, A., 1948, *J. Phys. Radium*, **9**, 88.
Single crystal of benzene.

FRÜHLING, A., 1949, *C.R. Acad. Sci., Paris*, **229**, 33.
Anthracene, low frequencies.

FRÜHLING, A., 1951, *Ann. Phys., Paris*, **6**, 401.
Benzene, diphenyl.

GROSS, E. F., KORSHUNOV, A. V., and SELKIN, V. A., 1950, *J. Exp. Theor. Phys., U.S.S.R.*, **20**, 292.
p-, m-, o-di-iodobenzenes: low frequencies.

ICHISHIMA, I., and MIZUSHIMA, S., 1950, *J. Chem. Phys.*, **18**, 1686.
Benzene.

KASTLER, A., and ROUSSET, A., 1941, *C. R. Acad. Sci., Paris*, **212**, 645.
Low-frequency lines.

KASTLER, A., and ROUSSET, A., 1947, *Phys. Rev.*, **71**, 455.
Origin of low-frequency lines.

KORSHUNOV, A. V., 1950, *Dokl. Akad. Nauk, S.S.S.R.*, **74**, 691.
p-dichlorbenzene, p-brom-chlorbenzene, p-dibrom-benzene.

KORSHUNOV, A. V., and SELKIN, V. A., 1950, *J. Tech. Phys., U.S.S.R.*, **20**, 745.
o-di-iodobenzene : strong illumination apparatus.
MAJUMDAR, N. C., 1949, *Ind. J. Phys.*, **23**, 465.
CH_3Cl and CH_2Cl_2 : change of state.
MARIGNAN, R., and BARDET, L., 1952 a, *C. R. Acad. Sci., Paris*, **234**, 714.
Oxalic acid.
MARIGNAN, R., and BARDET, L., 1952 b, *C. R. Acad. Sci., Paris*, **234**, 1964.
Oxalic acid dihydrate.
MIZUSHIMA, S., MORINO, Y., SIMANOUTI, T., and KURATANI, K., 1949, *J. Chem. Phys.*, **17**, 838.
Hexachlorethane : change of state, structure.
MIZUSHIMA, S., SHIMANOUCHI, T., MIYAZAWA, T., ABE, K., and YASUMI, M., 1951, *J. Chem. Phys.*, **19**, 1477.
Ethylene chlorhydrin.
MUKERJI, S. K., and LAL, B., 1951 a, *Ind. J. Phys.*, **25**, 53.
Diphenylmethane.
MUKERJI, S. K., and LAL, B., 1951 b, *Ind. J. Phys.*, **25**, 309.
Thianthrene.
MUKERJI, S. K., and SINGH, L., 1946, *Phil. Mag.*, **37**, 874.
p-diphenyl benzene : plane structure.
NARAIN, H., and SAKSENA, B. D., 1951, *Ind. J. Phys.*, **25**, 79.
p-dichlorbenzene : liquid and different orientations of solid.
PADMANABHAN, V. M., 1948 a, *Proc. Ind. Acad. Sci.* A, **28**, 485.
Na and K tartrates.
PADAMANABHAN, V. M., 1948 b, *Proc. Ind. Acad. Sci.* A, **28**, 494.
Ethylene diamine tartrate.
PADAMANABHAN, V. M., 1950, *Proc. Ind. Acad. Sci.* A, **32**, 184.
Sodium ammonium tartrate.
PADAMANABHAN, V. M., 1951, *Proc. Ind. Acad. Sci.* A, **34**, 155.
Na and K tartrates.
PALM, A., 1951, *J. Phys. Coll. Chem.*, **55**, 1320.
Polystyrene.
PATEL, M. M., 1949, *Curr. Sci.*, **18**, 436.
Methyl methacrylate.
RASKIN, S. S., and SKRIPOV, F. I., 1949, *Dokl. Akad. Nauk, S.S.S.R.*, **64**, 317.
d-tartaric acid : translational frequencies.
RAY, A. K., 1951, *Ind. J. Phys.*, **25**, 459.
p-dichlorbenzene.
ROUSSET, A., 1945, *Ann. Phys., Paris*, **20**, 53.
Rotational oscillations of molecules in organic crystals.
ROUSSET, A., 1947, *La diffusion de la lumière par les molécules rigides* (Paris: C.N.R.S.).
Monograph.
ROUSSET, A., 1948, *J. Phys. Radium*, **9**, 101.
Rotational oscillations and thermal agitation.
SAKSENA, B. D., 1950, *J. Chem. Phys.*, **18**, 1653.
p-dichlorbenzene : effect of rapid cooling.
SANYAL, S. B., 1950, *Ind. J. Phys.*, **24**, 378.
Monosubstituted n-butanes.
SANYAL, S. B., and MAJUMDAR, M. M., 1949, *Ind. J. Phys.*, **23**, 1.
Chloro- and bromo-acetyl chlorides : change of state.
SIRKAR, S. C., 1951, *J. Chem. Phys.*, **19**, 256.
Benzene : nature of low-frequency vibrations.
WEIL, A., and MATHIEU, J. P., 1952, *C.R. Acad. Sci., Paris*, **234**, 1868.
d-tartaric acid.

PARAMAGNETIC RESONANCE

By B. BLEANEY and K. W. H. STEVENS
Clarendon Laboratory, Oxford

CONTENTS

	PAGE
§ 1. Foreword	109
§ 2. Introduction	110
§ 3. The magnetic resonance phenomenon	112
3.1. The resonance condition	112
3.2. The fine structure	113
3.3. The hyperfine structure	115
§ 4. Line width	116
4.1. General discussion	116
4.2. Spin–lattice interaction	117
4.3. Spin–spin interaction	117
4.4. Experimental aspects	118
§ 5. Experimental techniques	119
5.1. General discussion	119
5.2. Ultimate sensitivity	121
5.3. The resonant cavity	122
5.4. The spectrometer system	123
5.5. Methods using modulation of the magnetic field	125
§ 6. Theory	125
6.1. The Hamiltonian	126
6.2. The crystal field for the iron group	128
6.3. Matrix elements of the crystal field	129
§ 7. The spin-Hamiltonian	131
7.1. Derivation	131
7.2. General discussion	134
7.3. S-state ions	136
7.4. The anomalous hyperfine structure	138
7.5. The resonance spectrum	139
§ 8. The iron group	140
8.1. Experimental results	140
8.2. Covalent bonding	146
§ 9. The rare earth group	147
9.1. Theory	147
9.2. Experimental results	149
§ 10. Exchange interactions	151
10.1. Isotropic and anisotropic exchange	151
10.2. Line widths	151
10.3. Copper sulphate and acetate	152
§ 11. Resonance in gases and other special compounds	154
11.1. Paramagnetic gases	154
11.2. Free radicals, phosphors, etc	156
§ 12. Conclusion	157
References	157

ABSTRACT. After a brief introduction a general account is given of the main phenomena associated with electronic magnetic resonance. Two chapters on line width and experimental technique review the problems involved in detection of the resonance phenomena. A short account of the theory of paramagnetism of the solid state follows, leading to a derivation of the spin-Hamiltonian. The importance of the latter as a method of describing the resonance results is emphasized, the discussion proceeding in terms which are, as far as possible, independent of the underlying theory. The ions of the iron (3d) group and the rare earth (4f) group are then surveyed individually in terms of the spin-Hamiltonian. The influence of exchange interaction is treated in a separate chapter, and the review ends with a brief account of resonance in gases and other special compounds.

§1. FOREWORD

THE properties of paramagnetic salts have been investigated by a number of methods, chief of which are measurement of the susceptibility and the paramagnetic rotation, together with the radio-frequency methods of paramagnetic relaxation and resonance. In a general account of paramagnetism it would be necessary to consider all of these, and to show how the experimental results from each method can be related to the theory. Such an account would be outside the scope of a Report, and would amount to a volume of considerable size. Moreover, adequate reviews of one method, paramagnetic relaxation, already exist (Gorter 1947, Cooke 1950). In the present Report it has therefore been decided to deal only with one aspect of paramagnetism, the resonance phenomenon. Magnetic resonance forms the basis of a branch of spectroscopy which has been a valuable addition to the methods of investigating the solid state. As such, it may broadly be divided into four aspects—nuclear, ferromagnetic, antiferromagnetic and paramagnetic resonance. The first of these is concerned with nuclear dipoles, the others with electronic dipoles. Ferro- and antiferromagnetic resonance deal with magnetic systems where the electronic dipoles are strongly coupled together by exchange forces. Paramagnetic resonance, with which alone we shall be concerned in this Report, is confined to loosely coupled systems, where the paramagnetic units may be regarded as individuals.

The properties of such 'magnetically dilute' systems are of interest not only for their own sake, but also because of the role they play in low temperature physics. For the attainment of temperatures below 1° K they constitute the 'working substance', but before the application of radio-frequency methods the properties which determine their usefulness for this purpose could only be adequately determined by measurements in the region below 1° K. Such measurements are difficult, and expensive in man-hours. The resonance method of investigation is quicker, since the behaviour of the energy levels can be studied at higher temperatures. By the use of very dilute salts, complex systems of energy levels, with separations as small as 10^{-3} deg K can be analysed. This is especially valuable in salts where the nucleus of the paramagnetic ion gives rise to a hyperfine structure, and the results of resonance measurements have played a marked role in the successful development of methods of nuclear alignment.

No report on paramagnetism would be complete without some account of the theory of the crystalline electric field. Details of this are beyond the scope of the average experimentalist, but for many purposes, the magnetic properties of a dilute salt can be adequately described by the 'spin-Hamiltonian', a

derivation and discussion of which are presented in §7. We have endeavoured to make this discussion adequate for an understanding of the 'spin-Hamiltonian' without reference to the crystalline field theory, and we hope that this arrangement will facilitate the use of this Report by those for whom paramagnetism is a means rather than an end.

§2. INTRODUCTION

The distinction between paramagnetic and diamagnetic substances has been known for a very long time, but it may fairly be said that the foundations of the modern theory of paramagnetism rest on the work of Pierre Curie, who discovered experimentally the law which bears his name, and the theoretical deduction of this law by Langevin. This work showed that the distinction between diamagnetic substances and paramagnetic substances arose from the presence of permanent magnetic dipoles in the latter. The further introduction by Weiss of a strong internal coupling between these permanent dipoles gave a qualitative explanation of ferromagnetism. In this report we shall be concerned only with paramagnetism; that is, the magnetic properties of substances where each permanent magnetic dipole is substantially independent of its neighbours. The theory then treats each dipole separately, and the effects of neighbouring dipoles enter only in a subsidiary manner.

The presence of permanent magnetic dipoles received a natural explanation with the introduction of quantum theory, and they occur whenever an atom or an ion contains a partly filled electron shell. Such systems have a resultant angular momentum due partly to the orbital motion and partly to the intrinsic spin of the electrons, with each of which is associated a magnetic dipole moment. The requirement of a partly filled electron shell restricts the occurrence of paramagnetism in compounds to certain well-defined regions of the periodic table, known as the transition groups. Of these the most fully investigated are the iron group, where the 3d shell is incomplete, and the rare earth group (4f shell). Much less is known about the palladium (4d), platinum (5d) and uranium (5f–6d) groups.

The quantitative theory of paramagnetism was placed on a firm basis only after the advent of wave mechanics, which removed certain difficulties encountered in the older quantum theory (see Van Vleck 1932). Most paramagnetic compounds are solids, and in the iron group there remained the difficulty that the susceptibility was much closer to that which would be expected from the electron spins alone than that calculated for an agglomeration of free ions. The reason for this was found to lie in the strong interaction, in the solid state, between the paramagnetic ion and its immediate, diamagnetic neighbours. These are charged and produce a strong electrostatic field, through which the electrons of the paramagnetic ion must move. The interaction between the electrons and this field is so great that in the iron group, where the paramagnetic 3d electrons are the outermost electrons and fully exposed to the field of the neighbours, the orbital motion is 'quenched'. This means that the orbital motion is 'locked into' the field of the neighbours, and is unable to make its proper contribution to the magnetism. The electron spin, however, with its corresponding magnetic moment, has no direct interaction with the electrostatic field, and remains free to orient itself in an external magnetic field. Thus the magnetic properties of salts of the iron group correspond substantially to

'spin-only' magnetism. In the rare earth transition group, the unfilled electron shell is an inner 4f shell, and the interaction with the electric field of neighbouring ions ('the crystalline electric field') is much less. The paramagnetism of these ions is therefore, in a first approximation, close to that of free ions, both orbit and spin making contributions.

The bulk of the experimental evidence on which the theory of the crystalline electric field interaction was based came from measurements of the susceptibility, including the study of anisotropy in single crystals. This anisotropy arises as follows. In a free ion the motion of the electrons in their orbits sets up a magnetic field which interacts with the spin moments, and an external magnetic field orients both the spins and the orbits. In the crystal field the orbits may be greatly distorted and the associated magnetic field, which is proportional to the resultant orbital moment, will have directional properties determined by the crystal field. It may thus happen that the orbital moment is much reduced and also will not be oriented so readily by an external magnetic field. The spin moment sets itself along the resultant of the two fields and anisotropy results. When the orbital moment is very small, the internal field is small, and spin-only paramagnetism results, with small anisotropy. When the orbital motion is only partially quenched there is some orbital moment remaining, the internal field is appreciable and is to some extent tied to the crystal field, and there is large magnetic anisotropy.

In quantum mechanical language these effects may be stated as follows. When the interaction of the orbital motion with the crystalline field is much stronger than that between the spin and the orbits (the spin–orbit coupling) the spatial degeneracy is partially lifted (in the iron group the overall splitting may amount to 20 000 cm^{-1}). A singlet orbital state has zero magnetic moment, so that when a singlet is lowest and the spin–orbit coupling is neglected, the magnetism comes entirely from the spin. Inclusion of the spin–orbit coupling results in there being admixtures of other orbital states which are small provided the separations to these states are large compared with the spin–orbit coupling. In this case the susceptibility approximates to that of the spin only, and follows Curie's law down to very low temperatures with small anisotropy. When the ground orbital state is not a singlet, or is a singlet with another orbital level near, the orbital magnetism is only partially quenched and is tied to the crystal field. The susceptibility departs considerably from the spin-only value and from Curie's law and is strongly anisotropic.

From an analysis of the major terms in the crystalline electric field, Van Vleck (1932) was able to account quantitatively for the behaviour of the susceptibility of various ions in the 3d group. Schlapp and Penney (1932) applied this theory in detail to a number of salts, and more recent work has led to further developments of Van Vleck's ideas. The progress of the theory was hampered by lack of precise experimental data. These data came mainly from measurements of susceptibility, usually of powders but in some instances of single crystals. These measurements were seldom carried out over a sufficiently wide range of temperatures and, in addition, the susceptibility is not a particularly favourable datum, even for a single crystal, for often the unit crystal cell contains several paramagnetic ions whose axes of anisotropy are differently oriented. The net anisotropy is then a sum of the individual contributions. Again, at temperatures such that a number of levels are populated, the susceptibility is

also an average over these levels. The particular advantage of the new paramagnetic resonance technique over these older methods arises from the precise data it gives for each ion, even though the data are usually restricted to the ground state.

§3. THE MAGNETIC RESONANCE PHENOMENON

3.1. *The Resonance Condition*

Radio-frequency magnetic resonance was first applied to the accurate determination of nuclear magnetic moments by Rabi, using the method of molecular beams. Later, the resonance method for nuclei was applied to aggregated matter by Purcell, Torrey and Pound (1946) and by Bloch, Hansen and Packard (1946). When applied to electronic paramagnetism, the principle of the method is the same, and can be outlined as follows: When an ion which possesses a resultant angular momentum, and a parallel permanent magnetic moment, is placed in a constant external magnetic field H,* the angular momentum vector precesses about the axis of H with the angular frequency

$$\omega = g(\boldsymbol{e}/2\boldsymbol{mc})H. \qquad \ldots\ldots(3.1)$$

Following Kittel (1949) we shall call g the spectroscopic splitting factor; it is unity if only orbital motion of the electrons exists, and the precession frequency is then that given by the classical theorem of Larmor. In general, however, g will be different from unity. If a circularly polarized magnetic field is also applied, with such a sense and frequency that it rotates about H in synchronism with the angular momentum vector (and hence with the magnetization vector), a constant couple acts on the latter causing it ultimately to turn over and reverse its projection on H. The ion has then exchanged energy with the radiation field. The quantum picture, for a free ion of resultant angular momentum J, is as follows. In the field H the levels corresponding to the various spatial orientations of J have energies equal to $Mg\beta H$, where M is the electronic magnetic quantum number, and β the Bohr magneton. The selection rules for the allowed transitions, magnetic dipole in character, are $\Delta M = \pm 1$, and can be induced by quanta of frequency such that

$$\boldsymbol{h}\nu = g\beta H. \qquad \ldots\ldots(3.2)$$

Insertion of the expression $(\boldsymbol{eh}/4\pi\boldsymbol{mc})$ for β shows that in this simple case the required frequency is identical with that given by (3.1). In a system of ions in thermal equilibrium with the surroundings, states of lower energy have the greater population. Since transitions up and down have equal *a priori* probability, the net result of the application of resonance radiation is a gain of energy from the radiation field, and a shift towards more equal population of the various levels. This corresponds, of course, to an increase in temperature of the system. The resonance absorption can be detected by the loss of energy from the radiation field, which causes a damping of the tuned circuit in which the paramagnetic substance is placed.

Experiments in paramagnetic resonance are generally carried out at centimetre wavelengths, using fields of a few kilogauss. It is obvious that measurement of the wavelength and the external field give, in the simple case assumed above, an immediate determination of g, and eqn. (3.2) may be rewritten

$$g = 21{\cdot}4178/H\lambda \qquad \ldots\ldots(3.3)$$

* Strictly speaking, the field should be written B, but in practice H is generally used. Whether this practice is redeemed, or made worse, by calling the unit the gauss, is a matter of opinion.

where H is measured in kilogauss and λ is the wavelength in centimetres. There are several reasons for the use of centimetre rather than longer wavelengths. The intensity of the lines is greater (see § 5), and the line width in normal paramagnetic salts is itself of the order of 0·01 to 0·1 cm^{-1}, so that no resolved lines would be observed except at higher wave numbers. In addition, a number of salts, such as those used in adiabatic demagnetization work, have initial splittings of the ground state of the order of 0·1 to 1 cm^{-1}. Use of wavelengths shorter than about 1 cm has been restricted partly by the lack of suitable sources and detectors of radiation, and partly by the large magnetic fields required.

The advantages of paramagnetic resonance may now be enumerated. In the main, anisotropy arises from variation of the spectroscopic splitting factor g as the orientation of the field H is changed with respect to the axes of the crystal field. In general, therefore, each type of ion in a salt will give its own spectrum, separate from those of other types. This means that impurities do not affect the results, a fact which is of great importance in substances such as the rare earths, which are often costly or not available in highly purified condition. The experimental results yield precise information for a given level, and the quantity of material required is very small (see § 5). The disadvantages are that single crystals are required for accurate work, and certain substances cannot be investigated because there are no allowed transitions between the levels which give rise to the paramagnetism.

3.2. *The Fine Structure*

In salts where the ground state is an orbital singlet with a multiplicity of more than 2 in the spin ($S=1$ or more), the degeneracy of the spin levels is often lifted even in the absence of an external magnetic field. Such salts are in common use in the method of adiabatic demagnetization, and the magnitude of these splittings has generally been found from measurements below 1°K of the associated specific heat anomaly. When magnetic dipole transitions are allowed between the separated levels, it is obvious that the splitting can also be found by the method of paramagnetic resonance, since absorption of energy will only occur when the frequency of the applied radio-frequency field is equal to that of the splitting. This however would involve measurement of the absolute absorption coefficient over a range of frequencies, which is difficult to make accurately.

Experimentally it is vastly more convenient to apply a magnetic field to vary the splittings and then examine the spectrum as a function of field at constant frequency. From this it is, in principle, possible to pass from high field measurements to the zero field splittings. An example will suffice to demonstrate this. We consider an ion of spin 3/2, with an initial splitting between the doublets $M=\pm\frac{1}{2}$ and $M=\pm 3/2$, due to an axial crystalline electric field. The axis of quantization is here taken as the same as the axis of the field. When a magnetic field is applied parallel to the axis the energy levels diverge linearly (see fig. 1), and the component of the oscillating magnetic field perpendicular to the axis induces transitions $\Delta M=\pm 1$, and the parallel component transitions $\Delta M=0$. The arrows of constant length correspond to a given quantum of energy and are drawn to indicate the $\Delta M=\pm 1$ transitions. It will be seen that these occur at different fields, giving a triplet structure to the

spectrum. It is convenient to borrow a common spectroscopic terminology and call this the 'fine structure' of the spectrum. It is clear that the zero field splittings are readily deduced from the fine structure. The fact that the levels diverge linearly is related to the fact that the axis of precession does not change when the field is applied. If however the field is applied at an angle to the axis there is competition between the crystal field and the magnetic field to determine the axis of precession. The behaviour of the energy levels is more complicated and the axis of precession changes from that of the crystal field to that of the magnetic field as the magnetic field increases. At high fields the lines diverge

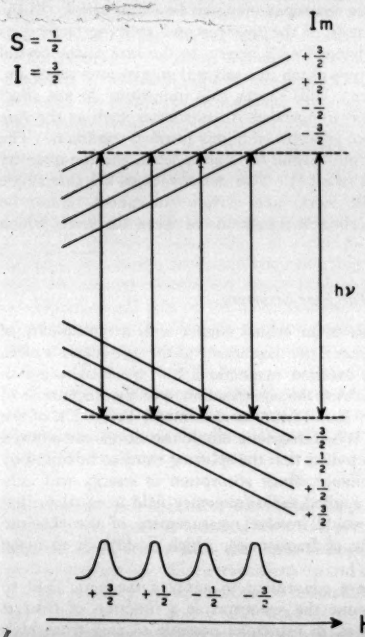


Fig. 3. Energy level diagram for $S=1/2$, $I=3/2$ in strong magnetic field. The allowed transitions are indicated (*after* Bleaney 1951 a). *Reproduced by courtesy of* '*Physica*'.

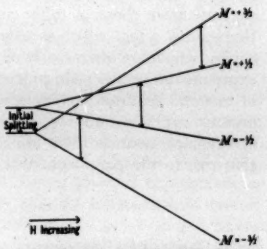


Fig. 1. Energy level diagram, $S=3/2$, magnetic field parallel to axis of crystalline field. The vertical arrows represent the allowed transitions at a given frequency. The value of g can be determined from the position of the central transition, and the splitting $2D$ from the separation between the outer transitions.

Fig. 2. Oscillogram showing fine structure of the trivalent gadolinium ion ($S=7/2$) in gadolinium ethyl sulphate (*after* H. E. D. Scovil).

linearly, giving again a fine structure which is independent of frequency. It may be remarked that with the oscillating and steady magnetic fields mutually perpendicular the lines are of unequal intensities, the central lines being stronger since they are due to transitions between states for which the projection of the magnetic moment on the direction of the radio-frequency magnetic field is greatest. A typical fine structure is shown in fig. 2.

In the above discussion we have chosen the crystalline field axis as axis of quantization. At high magnetic fields it is more natural to use the direction of H

but, whichever is used, at intermediate fields the actual states will, in general, be admixtures of states of different M. The selection rules are still $\Delta M=0$ for oscillating fields parallel to the axis of quantization and $\Delta M=\pm 1$ for perpendicular fields, so transitions may occur between any states which contain values of M differing by $0, \pm 1$. If for simplicity the levels are labelled by their strong field, or zero field states, these transitions may correspond to $\Delta M=\pm 2$, or more, which are sometimes referred to as forbidden transitions (Bleaney 1950b, Ubbink, Poulis and Gorter 1951, Malvano and Panetti 1950). The intensity of all such forbidden transitions becomes vanishingly small as the field is increased.

3.3. *The Hyperfine Structure*

When the nucleus of the paramagnetic ion also possesses a resultant angular momentum and hence a nuclear magnetic moment, a hyperfine structure is observed in the resonance spectrum. If the nuclear spin is I, then in strong external magnetic fields each electronic transition is split into $2I+1$ equally spaced lines of equal intensity. The reason for this simple splitting is as follows. The nuclear orientation is quantized, and the projection of the nuclear moment has $2I+1$ allowed values, each proportional to the nuclear magnetic quantum number m, which takes all values differing by unity between $+I$ and $-I$. At a given frequency, the electron will make a transition when the net field in which it finds itself corresponds to that of eqn. (3.2). This net field has $2I+1$ equally spaced values (each corresponding to a particular nuclear orientation), and there is a corresponding splitting of the electronic transition.

In quantum mechanical language, the magnetic interactions between electrons and nucleus split each electronic level into $2I+1$ equally spaced levels, as shown in fig. 3 for a hypothetical case with $S=\frac{1}{2}$, $I=3/2$. The radio-frequency magnetic field exerts a negligible effect on the tiny nuclear moment so that the allowed transitions are those in which m does not change. These are shown by the arrows, all drawn with constant length to correspond to the constant frequency applied. The spectrum (displayed in field) consists therefore of four equally spaced lines. Since at normal temperatures all nuclear orientations have the same probability, the lines are of equal intensity, unlike those due to fine structure. The equal spacing is true only in strong external fields, when the energy levels vary linearly with the field. At lower frequencies, where the external field applied becomes comparable with that produced by the nucleus, the picture is more complicated. The levels do not vary linearly with the external field, and the hyperfine structure is no longer equally spaced. The states contain mixtures of the various nuclear magnetic quantum numbers m, and when two states each contain the same value of m a transition is allowed which appears to be 'forbidden' if we label the states by their strong field quantum numbers only.

Another type of 'forbidden transition' may be observed even in strong fields if, in addition to the magnetic interactions between electrons and nucleus, there is also an electrostatic interaction between the gradient of the electric field at the nucleus and the electric quadrupole moment of the nucleus. The effect of this quadrupole interaction is to change the energy of the hyperfine levels by an amount proportional to m^2, all levels with the same m being changed by the same amount. Thus the allowed transitions, for which $\Delta m=0$, are not affected (in the first approximation). When the external field is oriented so that the magnetic field of the electrons at the nucleus is parallel to the gradient of the

electric field, the system precesses about this axis, and the statement in the previous sentence is true in any approximation. When the directions of the magnetic field and the gradient of the electric field are not parallel the various nuclear states are admixed and forbidden transitions are allowed. The position of such lines in the spectrum depends on the nuclear quadrupole interaction, and affords a direct measure of the latter. If the quadrupole interaction is small such lines are weak (their intensity varies as the square of the interaction) and may not be visible. An indication of the size of the quadrupole interaction can then be obtained from second-order shifts in the positions of the $\Delta m = 0$ lines, but this is not a very sensitive method.

Hyperfine structure in paramagnetic resonance was first observed by Penrose (1949), following a suggestion of Gorter (1948), and the quadrupole interaction was found by Ingram (1949). Since then hyperfine structures have been observed and analysed in many salts, but in most of them the quadrupole interaction is too small to be detected.

A survey of the experimental work on hyperfine structure has been given by Bleaney (1951 a). Various theoretical formulae for the positions of the allowed transitions can be found in Bleaney (1951 b), and fig. 4 shows a typical case.

Fig. 4. Oscillogram of hyperfine structure of the divalent cobalt ion ($S=1/2$, $I=7/2$) in cobalt bismuth nitrate. The unequal spacing is due to second order displacements which are appreciable in intermediate magnetic fields (*after* R. S. Trenam).

§4. LINE WIDTH

4.1. *General Discussion*

In microwave spectroscopy generally, resolution of the spectrum is limited, not by instrumental effects, but by the widths of the lines themselves. In paramagnetic resonance there are two major causes of line width—interaction between the paramagnetic ions and the lattice, and interaction between the various ions themselves. These interactions have often been treated as giving relaxation effects, characterized by two relaxation times, the spin–lattice relaxation time τ and the spin–spin relaxation time. The extensive investigation of relaxation effects by Gorter (1947) and others, at frequencies of the order of the reciprocal of the relaxation time, has already been the subject of a report by A. H. Cooke (1950). Such frequencies do not give resolved lines, since a relaxation process gives a half width at half intensity of the order of $1/2\pi\tau$ sec^{-1}. In paramagnetic resonance work resolved lines are required and higher frequencies must be used. One prime objective has been the reduction of line widths in order to increase the resolution. A second objective has been the study of the line width itself in order to obtain information about the interactions which contribute to it. These interactions will be discussed briefly in the following section. It will be seen that the spin–lattice interaction may be characterized by a relaxation time. The spin–spin interaction, on the other hand, is more complicated, and cannot fully be so described.

4.2. *Spin–Lattice Interaction*

The spin–lattice relaxation time (or more precisely, its inverse) is a measure of the rate at which a spin reverses direction and either takes from, or gives to, the lattice a quantum of energy. The theory of this is mainly due to Van Vleck (1940) and was given with particular reference to non-resonant absorption. No detailed application of these ideas has been given to resonant absorption, probably because the Van Vleck theory is not entirely satisfactory and in resonance absorption it is usual to try to eliminate the effects of spin–lattice relaxation by working at low temperatures. The mechanism considered is one in which the thermal vibrations of the lattice give a fluctuating crystalline electric field. In ions where the magnetism comes mainly from the spin, on which the electric field has no direct action, the spin–orbit coupling plays an essential role as the mechanism by which the spin feels the effect of the thermal vibrations. The development of the theory shows that the spin–lattice relaxation time τ is strongly temperature dependent, becoming longer as the temperature is reduced. The theory also shows that τ depends very markedly on the separation between the ground state and the first excited state—if this separation is large τ is long, and if the separation is small τ is short. These features are observed experimentally, though the actual dependence on the separation and on temperature do not seem to follow the expressions given by the theory (Bagguley and Griffiths 1952). It may also be remarked that the other effects of thermal vibrations, such as modulation of the dipole–dipole interactions between the spins, appear to give negligible widths.

4.3. *Spin–Spin Interaction*

The second broadening process arises from the interactions between the dipoles, which can be regarded as situated in a rigid lattice. This broadening shows no dependence on temperature (above the Curie point) and can only be reduced by separating the magnetic carriers. Two main types of interactions between the ions have been recognized, the dipole–dipole and the exchange interactions. The exchange will be considered separately in §10. The dipole–dipole interaction is that due to the magnetic moments of the ions, which can be regarded as rather like bar magnets. As with all interactions of this sort, it depends on the angle between the spins and on the angle between the spin and the vector to the other ion. A picture of the broadening process can be obtained as follows. Each bar magnet is regarded as precessing about the external field, and can be resolved into a component which is steady and directed along the field, together with a rotating component at right angles. The steady component sets up a steady field at any other magnet, which therefore behaves as if it is in a field which is slightly different from the external field. The rotating component sets up a rotating field, and if this rotating field has the same frequency as the frequency of precession of some other magnet there will be a couple acting on this second magnet tending to change its direction. The first process gives a broadening which is rather like that which arises from using an inhomogeneous magnetic field, and the second gives a resonance broadening because it tends to reduce the lifetime of an ion in a given state. The theory of spin–spin interactions has been developed by Van Vleck (1948) and Pryce and Stevens (1950). In giving an account of this it is necessary to include exchange interactions and it

is therefore convenient to postpone the discussion to §10, where exchange is considered in its wider aspects. For convenience, however, we quote the formula for the mean square width of a line, given by Van Vleck for the case of free spins, assuming dipole–dipole interactions only

$$\hbar^2(\Delta\nu^2)_{\mathrm{av}} = \tfrac{3}{4}S(S+1)g^4\beta^4\Sigma_k\left(\frac{1-3\cos^2\theta}{r^3}\right)^2_{jk} + \tfrac{1}{3}S'(S'+1)g^2g'^2\beta^4\Sigma_{k'}\left(\frac{1-3\cos^2\theta}{r^3}\right)^2_{jk'}$$

where the resonance is from ions of type j (spin S, spectroscopic splitting factor g), the dissimilar ions are of type k' (S'; g') and θ denotes the angle between the line joining two ions and the direction of the external field. It will be noticed that similar ions are $3/2 = (\frac{3}{4}/\frac{1}{3})^{1/2}$ times as effective as dissimilar ions in broadening the line, thus showing the effect of resonance broadening.

4.4. *Experimental Aspects*

The great variation in spin–lattice relaxation times is illustrated by the fact that τ ranges at 10°K from about 10^{-4} second to less than 10^{-11} second. The latter represents the limit of observation, since the line width is then of the order of 1 cm^{-1} and comparable with the frequencies used in resonance work. In a number of compounds the relaxation time is sufficiently long for resonance lines to be observed at room temperature, but this is not true of others. Fortunately, the rapid increase in τ when the temperature is lowered means that most salts of the iron group can be investigated at the convenient temperatures provided by liquid oxygen, nitrogen and hydrogen. In other cases liquid helium is required. In general, relaxation times longer than 10^{-7} second give a contribution to line width which is negligible compared with that from spin–spin interaction. If τ is longer than about 10^{-5} second, however, undesirable saturation effects appear, as the mechanism for transferring heat from the spin-system to the lattice is unable to cope with the heat absorbed by the spin-system from the radiation field. In consequence the temperature of the spin-system rises, the intensity of the spectrum falls, and the line shape may be considerably distorted.

The resonance method has not been greatly used for the determination of spin–lattice relaxation times, though the region ($\tau < 10^{-7}$ sec) in which it is immediately applicable is complementary to that covered by the method of paramagnetic relaxation ($\tau > 10^{-7}$ sec). When spin–lattice relaxation is the major source of line width, the line shape is similar to that of collision-broadened lines, and the formula of Van Vleck and Weisskopf (1945) is valid. The relaxation time may thus be found directly from the half-width. Schlichter and Purcell (1949) have reported a determination of τ, when it is longer than 10^{-7} second, by use of the saturation phenomenon, and this method is valuable for salts of great dilution where other methods are not available (Schneider and England 1951).

Unlike spin–lattice interaction, spin–spin interaction is substantially independent of temperature and can only be reduced by increasing the distance between the paramagnetic ions. For this reason complex salts are generally used, where the paramagnetic ions are imbedded in, and well separated by, a mass of diamagnetic ions. For example, in $CuK_2(SO_4)_2.6H_2O$, the only paramagnetic ion is the Cu^{2+} ion, and the least distance between two such ions is 6 Å. With a magnetic moment of one Bohr magneton, the field of a nearest ion

is still about 50 gauss, and the sum over the nearest neighbours amounts to one or two hundred gauss. If one or two neighbours are substantially closer than any others, the line width due to magnetic dipole interaction is anisotropic, owing to the directional properties of the magnetic interaction. For two magnets this varies as $|1-3\cos^2\theta|$, where θ is the angle between the radius vector joining the two and the direction of the external field. A variation which approximates to this is found in the copper Tutton salts (Bleaney, Penrose and Plumpton 1949). In extreme cases of one or two close neighbours, the resonance line has a resolved structure (Bleaney, Elliott and Scovil 1951), satellites of the central line appearing at points on either side displaced by an amount equal to the combined fields of the neighbours. In general, of course, a broadened line is made up of a large number of such satellites, and its shape depends on the arrangement of the nearest neighbours. If these are more or less symmetrically distributed, the line shape often approximates to that of a gaussian error distribution, as in caesium chrome alum (Bleaney 1950b).

To reduce the line width further, the paramagnetic salt may be 'diluted' by growing it as a mixed crystal together with a diamagnetic substance. Thus, the salt $CuK_2(SO_4)_2.6H_2O$ mentioned above may be diluted by replacing the copper by zinc or magnesium, both of which are diamagnetic and form crystals isomorphous with the copper salt. Dilution narrows the lines because the number of neighbours in a given position which are paramagnetic is reduced in proportion to the dilution. As the latter is increased it is found that a limiting width is attained, of the order of 6 gauss half width for salts such as that just given. This width arises from the random fields of the nuclear magnetic moments of the protons in the waters of crystallization. These are generally located immediately around the paramagnetic ion (at about 2 Å distance), and are not, of course, removed by dilution. It is possible, however, to reduce the half width to about 2 gauss by growing crystals from heavy water (Bleaney, Bowers and Ingram 1951), since the magnetic moment of the deuteron is only one third that of the proton. The latter is the chief impediment of further resolution, and it would be preferable to use substances without water of crystallization. These are not readily grown as single crystals, however, except from the melt. An exception is formed by the complex cyanides, such as $K_3Cr(CN)_6$, which can be grown from aqueous solution. Here the paramagnetic ion (Cr^{3+} in the quoted example) is surrounded by the six cyanides, the carbons (with no nuclear moment) being nearest, and the nitrogens (with a small moment) being next. Bowers (1952) has obtained lines with half widths of only 1 gauss in diluted salts of this compound.

§5. EXPERIMENTAL TECHNIQUES

5.1. *General Discussion*

The general technique of measurements at centimetre wavelengths is well known, and its application to the spectroscopy of gases has already been the subject of two review articles (Bleaney 1946–7, Gordy 1948). The technique applied to paramagnetic resonance is very similar, the only major difference being that the sample is placed in a resonant cavity rather than in an extended waveguide. The reasons for this are (*a*) the sample must be located in a strong homogeneous magnetic field, (*b*) the samples are usually small single crystals, and the highest sensitivity is obtained by concentrating the microwave field in

them. By so doing saturation effects are increased, but this is not as serious as in gaseous spectroscopy because the transition probabilities associated with magnetic dipoles ($\sim 10^{-20}$ e.m.u.) are much smaller than those for electric dipoles ($\sim 10^{-18}$ e.s.u.).

If the sample is subjected to an oscillating magnetic field $H_1 \cos \omega t$, the power absorbed by it is $\frac{1}{4}\omega\chi'' H_1^2$ where χ'' is the imaginary part of the susceptibility of the sample. This must be compared with the power absorbed in the resonant cavity itself, which may be written

$$\frac{1}{Q_0}\,\omega\,\frac{H_1^2}{8\pi}\,V$$

where Q_0 is the 'quality factor' of the unloaded cavity. Here V is the 'effective volume' of the cavity, defined by the equation $\int H^2\,dv = H_1^2 V$. Thus $(H_1^2/8\pi)V$ is the microwave energy stored in the cavity. The important quantity as far as detection is concerned is the ratio of the power absorbed in the sample to that dissipated in the cavity. From the above equations this is $4\pi\chi'' Q_0/V$, which is also the fractional change in the resistance due to the magnetic absorption. It is obvious that the greatest sensitivity will be obtained by making Q_0 as high as possible, and by keeping the effective volume of the cavity as small as possible. The latter is achieved, partly by making the actual volume small, and partly by locating the sample at the point of highest radio-frequency magnetic field strength. For a typical cavity at 3 cm wavelength $Q_0 = 5000$, $V = 3$ cm^3, and the power absorbed in the crystal will equal that dissipated in the resonator (i.e. the Q of the cavity will be halved) by a sample with a χ'' of about 5×10^{-5}.

For a sample containing N paramagnetic ions of spin S, the value of χ'' for the $M \longleftrightarrow M-1$ transition (where M is the electronic magnetic quantum number) is

$$\chi'' = \frac{\pi\nu g'^2\beta^2 N\{S(S+1)-M(M-1)\}}{8kT(2S+1)}\,f(\nu) \qquad \ldots\ldots(5.1)$$

where ν is the frequency at which resonance is observed, and $f(\nu)$ is a function depending on the line shape, which is normalized so that its total value integrated over any one line is unity. At the centre of a line $f(\nu)$ is approximately equal to $1/\Delta\nu$, where $\Delta\nu$ is the half width at half intensity (the equality is exact for a line shape such as that observed for collision-broadened lines in gases). β is the Bohr magneton, and g' is a spectroscopic splitting factor which equals the ordinary g-value if the latter is isotropic, but otherwise depends on the orientation of the radio-frequency magnetic field relative both to the external field and to the crystal axes. For the special case of axial anisotropy, with the external field at an angle θ with the anisotropy axis, and the radio-frequency field perpendicular to the external field,

$$g' = g_\parallel g_\perp (g_\parallel^2 \cos^2\theta + g_\perp^2 \sin^2\theta)^{-1/2}. \qquad \ldots\ldots(5.2)$$

If the external and radio-frequency magnetic fields are not mutually perpendicular, g' should be multiplied by the sine of the angle between them. For the simple case of $S = \frac{1}{2}$, with an isotropic g-value, χ'' is approximately equal to $\chi_0(\nu/\Delta\nu)$, where χ_0 is the isothermal susceptibility of the sample, and χ'' is therefore considerably higher than χ_0 for well resolved lines. For a mole of paramagnetic ions with $S = \frac{1}{2}$, and $g = 2$, $\chi_0 = 0{\cdot}38/T$. If the ions are in a salt whose magnetic dilution is about that of the Tutton salts (ionic volume about 200 cm^3), the line width is about 100 gauss $= 0{\cdot}01$ cm^{-1} (assuming it is due to magnetic dipole

interaction alone), and hence at 3 cm wavelength ($\bar{\nu}=0{\cdot}3\,\text{cm}^{-1}$) χ'' would be roughly 4×10^{-2} per mole at room temperature, and in our example the Q of the cavity would be halved by a crystal of volume just over 0·1 cm³. These figures hold roughly also for the individual lines of the fine structure of ions with S greater than $\frac{1}{2}$, since χ_0 increases as $S(S+1)$, $\Delta\nu$ increases as $\{S(S+1)\}^{1/2}$, and there are $2S$ lines.

5.2. *Ultimate Sensitivity*

To evaluate the limiting sensitivity we shall assume a simple equivalent circuit, where the resonant cavity is represented by a tuned circuit of series resistance r. Microwave radiation is coupled into this through a mutual inductance M from an oscillator of available power P_1, of internal impedance R. The radiation is detected by a receiver also of impedance R, coupled to the tuned circuit by a second mutual inductance M. If the oscillator has angular frequency ω, and is exactly tuned to the resonant circuit, the voltage across the detector is

$$V_2=r'(RP_1)^{1/2}/(r+r') \qquad \ldots\ldots(5.3)$$

where $r'=2\omega^2M^2/R$. If, owing to the magnetic absorption, the series resistance changes by δr, the change in the detector voltage is

$$\delta V_2=r'(RP_1)^{1/2}\delta r/(r+r')^2. \qquad \ldots\ldots(5.4)$$

Since r' depends on the tightness of coupling into the resonant circuit, we can adjust it by changing the coupling (i.e. the value of M). The sensitivity will be greatest when the voltage δV_2 at the detector is a maximum, which occurs when $r'=r$. The way in which an optimum arises can be seen by re-writing (5.4) in terms of the power P_2 reaching the receiver. We have then

$$\delta V_2=\delta r(RP_2)^{1/2}/(r+r') \qquad \ldots\ldots(5.5)$$

which shows that the fractional change $\delta V_2/(RP_2)^{1/2}$ in the voltage at the receiver is equal to the fractional change in the 'loaded Q', which is $\delta r/(r+r')$. Thus, as we increase the coupling, P_2 goes up but this is finally less important than the fall in the loaded Q due to the heavier coupling. At the optimum, the loaded Q is just half the unloaded Q, and the detector voltage change due to the magnetic absorption becomes $\delta V_2=(RP_1)^{1/2}(\delta r/4r)$. If the receiver has a noise factor N and band width df, then the signal output will be equal to the noise output when $\delta V_2=(N\mathbf{k}T\,df R)^{1/2}$. Also $\delta r/r$ represents the ratio of the power absorbed by the sample to that dissipated in the resonator, i.e. $\delta r/r=4\pi\chi''Q_0/V$. Combining these two expressions, we find for the minimum value of χ'' the equation

$$\chi''_{\min}=\frac{V}{\pi Q_0}\left(\frac{N\mathbf{k}T\,df}{P_1}\right)^{1/2}. \qquad \ldots\ldots(5.6)$$

With the resonant cavity assumed previously, at 3 cm wavelength, a receiver noise factor of 10, a band width of 1 c/s, and an oscillator power of 40 milliwatts, the theoretical minimum value of χ'' is 10^{-13}. With a diluted salt the line width can be reduced so that $\nu/\Delta\nu$ is about 10^3, so that this value of χ'' corresponds to a χ_0 of about 2×10^{-16}, which would be achieved at 20° K with 10^{-14} moles of an ion of $S=\frac{1}{2}$, $g=2$. If the gram-ionic weight is about 100, the mass required is about 10^{-12} grammes.

One or two points in this analysis need a word of comment. If the sensitivity is expressed as a susceptibility per unit volume (χ''/V), as would be appropriate in dealing with a gas which fills the resonator, then the limiting value $\chi''_{\min}/V$

increases as the frequency rises, since on the whole Q_0 and P_1 are diminishing, while N is increasing. For a given density of gas, these changes are roughly compensated by the fact that the value of χ''/V produced by the resonance phenomenon also rises with the frequency (see eqn. (5.1)). On the other hand, if the total number of paramagnetic ions is fixed, as might be the case for a radioactive isotope, and they are in a solid sample where the line width is fixed by interaction with other atoms, it is much more favourable to use short wavelengths owing to the better filling factor; that is, the smaller value of the effective volume V in eqn. (5.6).

5.3. *The Resonant Cavity*

A cavity resonator design which has been much used in the Clarendon Laboratory for work at 90° K and 20° K is shown in fig. 5. The cavity resonates at about 3·2 cm wavelength, and consists of a rectangular H_1 waveguide, half a wavelength long. The cavity is in two halves, the join occurring at a point where no radio-frequency currents are interrupted, and the two halves are affixed by a coned ground joint external to the cavity. For this purpose the

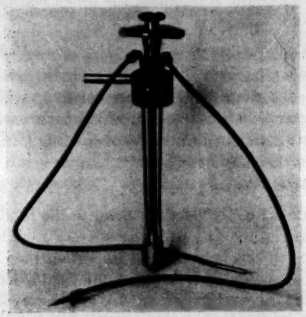


Fig. 5. Cavity resonator for 3 cm wavelength and temperatures down to 10°K, in common use at Clarendon Laboratory, Oxford.

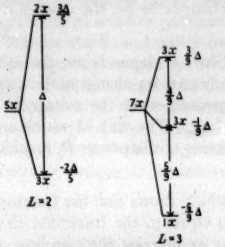


Fig. 6. Splitting of orbital levels by crystalline electric field of cubic symmetry for $L=2$ and $L=3$ respectively. The degeneracy of each level is indicated, x being the spin degeneracy.

external cross section is made circular. The ground joint is sealed with glycerine, which does not crack at low temperatures, and is vacuum tight. To reduce heat conduction, the microwave radiation is conveyed to and from the cavity by small coaxial lines, the outer conductor being formed by thin german silver tubing, the dielectric and inner conductor from flexible coaxial cable from which the braided outer has been stripped. The central conductor of this protrudes into the resonator to give a probe coupling, and a vacuum seal is made at the upper end where the cable emerges at room temperature, by using a rubber ring which is squeezed on to the dielectric.

In this design the paramagnetic crystal is mounted on the end of a thin german silver tube. This end is flush with the inside wall of the cavity, where the radio-frequency magnetic field is a maximum, when the mounting is inserted in the central tube shown in the figure. At the top the mounting terminates in a vacuum tight ground joint, surmounted by a dial graduated in degrees of angle.

By this arrangement the crystal can be rotated *in situ* while the apparatus is immersed in the refrigerant, and the spectrum for a whole plane of directions can be observed. The advantage of this system is that the external and radio-frequency magnetic fields (both horizontal) can be kept mutually perpendicular while the crystal is rotated, but it suffers from the disadvantage that rotation of a crystal with irregular shape detunes the cavity because of its rather high dielectric constant. This can be avoided if the crystal is fixed, and the whole assembly, or the magnet supplying the external field, is rotated instead about a vertical axis, but then the angle with the radio-frequency field varies and at some positions the intensity of the spectrum will become very small or vanish.

The assembly is immersed in about 200 cm^3 of refrigerant contained in a small glass Dewar vessel. With liquid hydrogen this lasts about half an hour, sufficient for most measurements. If the inside of the apparatus is filled with an atmosphere of hydrogen, the crystal mounting can be withdrawn and a new crystal inserted without removing the refrigerant.

The apparatus for the 1 cm wavelength region is similar in design, except that waveguide feeds (made from german silver tubing) are used instead of coaxial lines. Since the cavity is smaller, a thin-tailed Dewar vessel is used, which can be contained between magnet pole-pieces of 24 mm separation. This enables higher fields to be attained than with the separation of 42 mm required for the 3 cm apparatus.

For the wavelength region 6 to 20 cm a quarter-wave coaxial line resonator is used, the paramagnetic crystal being placed next to the centre conductor at the closed end. A series of resonators of different lengths is used to cover the region, each fitting into a vacuum jacket in the same apparatus. The coaxial line is mounted vertically, with the open end at the top, and over this moves a dielectric sleeve driven by a micrometer screw at room temperature. This enables the cavity to be tuned to intermediate wavelengths while immersed in the refrigerant.

For liquid helium temperatures, better heat insulation is required, and Bogle, Cooke and Whitley (1951) have used, at 3 cm wavelength, an apparatus similar to that in fig. 5, but with a much greater distance between the cavity and the top of the apparatus. The chief disadvantage of this is the increased 'microphony' as the apparatus is more sensitive to vibration. In an improved version all connections from room temperature go through a chamber containing liquid hydrogen, which thus intercepts most of the heat flow before it reaches the liquid helium. Weidner and Whitmer (1952 b) have used waveguide feeds formed by rectangular glass tubing internally silvered, which combines rigidity with small heat conduction.

5.4. *The Spectrometer System*

In the early work with concentrated crystals the magnetic absorption was sufficient to change the Q of the cavity appreciably and it could therefore be observed with a rather simple system. Power is fed from a microwave oscillator giving 20 to 100 milliwatts output through an attenuating waveguide or coaxial line to the cavity. Resonance in the latter is detected by using either the reflected power on the input line or the power flowing through a second output line coupled to the cavity. In the former case the magnetic absorption can be calculated from the formula $4\pi\chi'' Q_1/V = (|\gamma| - |\gamma_0|)/(1 - |\gamma|)$ where Q_1 is the 'loaded Q' of the cavity and γ_0 is the reflection coefficient when there is no magnetic absorption,

while γ is the reflection coefficient in the presence of magnetic absorption. In the second case $4\pi\chi''Q_1/V=(P_0/P)^{1/2}-1$ where P_0 and P are the powers flowing in the output line when magnetic absorption is absent and present respectively. In each case the oscillator must be exactly at the centre of the cavity resonance at every point of a magnetic resonance line. Since the latter introduces anomalous dispersion in the real part of the magnetic susceptibility, continual retuning is necessary, or distorted line shapes will be obtained.

With concentrated crystals, the line width is generally large and the absorption line can be recorded by a 'point-by-point' method. That is, the external magnetic field is adjusted to the desired value, the oscillator is tuned to resonance with the cavity, and the detector reading observed. With the transmission system, the power output from the cavity is fed to a silicon–tungsten crystal rectifier working in the square-law region, so that the rectified current read on a galvanometer is proportional to the power incident on the detector. In the reflection system the change in the reflection coefficient is determined by balancing the reflected and input powers with a 'magic tee' bridge in the absence of magnetic absorption, the unbalance which occurs due to the magnetic absorption being read again on a galvanometer fed by a silicon–tungsten crystal rectifier.

A much improved version of the single input system is used at Rutgers University (Weidner and Whitmer 1952 b). The oscillator is frequency modulated through a few hundred kilocycles per second at $\frac{1}{2}$ c/s so that it sweeps through the cavity resonance. This produces a periodic variation in the reflected power at 1 c/s. Since the reflected power is a minimum at the cavity resonance, the lower envelope of the trace it produces, after detection, on a recording galvanometer corresponds to pure absorption. Disturbing effects due to anomalous dispersion are thus avoided. The electromagnet produces a field which can be swept slowly over any desired range, and the system is specially designed for the accurate analysis of line shapes, either for very narrow lines, such as those of organic free radicals, or for wide absorption lines like those of the concentrated iron alums.

Bagguley and Griffiths (1952) have employed a balance system with two resonant cavities. The microwave oscillator is frequency modulated at 1 Mc/s, and this is converted to amplitude modulation at 2 Mc/s at the output from the resonators if they are tuned to the centre frequency of the oscillator. The signal from each resonator is rectified, and the 2 Mc/s modulation signals are fed in antiphase to a balanced transformer tuned to this frequency. A signal appears in the secondary of the transformer only when the two inputs are unbalanced through magnetic absorption in one of the resonators; this signal is amplified, and displayed on an oscilloscope. The external magnetic field could be swept through 4000 gauss in a period of from 10 to 20 seconds, and the x-sweep of the oscilloscope was synchronized with the field. The lines were photographed using recording paper.

Bagguley (unpublished) has developed another balance system, which is in principle analogous to the nuclear induction method of Bloch, Hansen and Packard (1946). A cylindrical H_{11} mode resonator is used and the steady magnetic field H_z is directed along the axis of the cavity. If H_x is the direction of the magnetic vector at one end of the cavity for the radio-frequency field exciting the resonator, the paramagnetic absorption is observed by measuring the induced field component H_y. (In practice it is more convenient to use probe coupling

to the cavity, so that actually the electric vector is used in detection.) Superheterodyne detection at 45 Mc/s is used and the voltage output of a 45 Mc/s amplifier is plotted as a function of the magnetic field H_z in order to delineate the absorption line.

5.5. *Methods using Modulation of the Magnetic Field*

If the external magnetic field is subjected to an amplitude modulation at a low frequency of p c/s, and during this modulation it sweeps across an absorption line, there will be a corresponding modulation of the power reflected or transmitted by the resonant cavity containing the paramagnetic sample. If this power is detected, the modulation can be amplified by a low-frequency receiver and displayed on an oscilloscope whose x-sweep is synchronized with the field modulation. This system has been used by Bleaney and Ingram (1949) and others, and the photographs of spectra in this report have been obtained by it. The field modulation frequency is 50 c/s, and the receiver has a flat band-pass from about 10 c/s to 20 000 c/s in order not to distort the line shape. The sensitivity is such that the signal from about a microgram of paramagnetic ions can be observed, and the chief limitation in sensitivity arises from the very high noise output of the silicon–tungsten crystal rectifier at audio frequencies.

Beringer and Castle (1950, 1951) have eliminated this source of noise by using a bolometer to detect the power transmitted through the cavity. The magnetic field is modulated at 30 c/s, the depth of modulation being small compared with the line width. The power reaching the bolometer is then modulated at 30 c/s to a depth which depends on the slope of the magnetic absorption line. The bolometer forms one arm of a high impedance d.c. bridge, adjusted so that the d.c. power in the bolometer equals the radio-frequency power (the most sensitive arrangement). A high gain audio-amplifier of 2 c/s band width, centred on 30 c/s, detects the out-of-balance signal from the bridge, which is then impressed on a 'lock-in' mixer whose d.c. output is read on a galvanometer of 0·04 c/s band width. With an improved 30 c/s pre-amplifier (Beringer and Castle 1951) the signal-to-noise ratio is unity for a gas which gives a χ'' (per unit volume) of 3×10^{-13}. This is probably the best yet achieved.

The second method of eliminating the abnormal crystal detector noise at audio frequencies is to balance down the microwave power by using a 'magic-tee' or similar bridge to a level at which a superheterodyne receiver and a local oscillator, followed by an intermediate-frequency amplifier at say, 45 Mc/s, can be used (Schneider and England 1951). Such a system permits the display of the spectrum on an oscilloscope and should be more sensitive by a factor of about 100 than that using an audio-frequency amplifier. The sensitivity can be further improved by using a narrow band width and small field modulation depth, with a correspondingly slow sweep of the magnetic field over an absorption line. The chief impediment then is the long period stability required of both microwave oscillator and external magnetic field. It is common to use a Pound frequency stabilizer to eliminate frequency fluctuations in the former.

§ 6. THEORY

In the previous discussion we have tried to give an outline of the principal features of the resonance phenomena. It is however just the fact that paramagnetic crystals do not behave as simply as the discussion might suggest

which makes the investigation of the phenomena such a fruitful source of experimental and theoretical problems. In the following sections our aim is to give an account of those aspects of the theory of the resonance which seem to be of most frequent occurrence. We begin by discussing the Hamiltonian which is used to describe the free ion, and will then consider the effect of adding the crystal field. This leads on to the problem of finding the matrix elements of the field and will be considered for the special case of a field of nearly cubic symmetry. In § 7 we shall give an account of the method of the spin-Hamiltonian and will discuss the problem of guessing a likely spin-Hamiltonian in the absence of a theory.

6.1. *The Hamiltonian*

We begin by enumerating the terms which occur in the Hamiltonian of the free ion, in order of decreasing magnitudes. They are:

(*a*) The coulomb interaction of the electrons with the nucleus (assumed fixed) and with each other

$$\sum_k \left(\frac{\mathbf{p}_k^2}{2m} - \frac{Ze^2}{r_k}\right) + \sum_{j<k} \frac{e^2}{r_{jk}}. \qquad \ldots\ldots(6.1)$$

(*b*) The magnetic interactions between the electron spins and the orbits (Van Vleck 1932)

$$\Sigma\{a_{jk}\mathbf{l}_j.\mathbf{s}_k + b_{jk}\mathbf{l}_j.\mathbf{l}_k + c_{jk}\mathbf{s}_j.\mathbf{s}_k\}. \qquad \ldots\ldots(6.2)$$

This form for the interaction is not the only one possible and Pryce (1950 a) has used a spin–spin interaction of the form

$$\Sigma 4\beta^2\left[\frac{\mathbf{s}_j.\mathbf{s}_k}{r_{jk}^3} - \frac{3(\mathbf{r}_{jk}.\mathbf{s}_j)(\mathbf{r}_{jk}.\mathbf{s}_k)}{r_{jk}^5}\right] \qquad \ldots\ldots(6.3)$$

to describe the interactions between the electron spins, based on the usual form for dipole–dipole coupling. The orders of magnitude are 10^2–10^3 cm^{-1} for the spin–orbit coupling and ~ 1 cm^{-1} for the spin–spin coupling.

(*c*) Interaction with an external magnetic field gives

$$\Sigma\frac{e\boldsymbol{h}}{4\pi mc}(\mathbf{l}+2\mathbf{s})_k.\mathbf{H} = \beta(\mathbf{L}+2\mathbf{S}).\mathbf{H} \qquad \ldots\ldots(6.4)$$

(where the term $(e^2H^2/8mc^2)\Sigma(x^2+y^2)_k$, which gives rise to diamagnetism, has been omitted as of no present interest). The associated energy is of order 1 cm^{-1}. It may be remarked that in very accurate work the gyromagnetic ratio of the electron spin should be taken as 2·0023 rather than 2. The remaining interactions are those which arise from couplings to the nucleus and are:

(*d*) The interaction between the magnetic moment of the nucleus and the magnetic field set up by the orbital and spin moments of the electrons (energy $\sim 10^{-2}$ cm^{-1})

$$2\gamma\beta\beta_N\Sigma_k\left[\left\{\frac{(\mathbf{l}_k-\mathbf{s}_k).\mathbf{I}}{r_k^3} + \frac{3(\mathbf{r}_k.\mathbf{s}_k).(\mathbf{r}_k.\mathbf{I})}{r_k^5}\right\} + \frac{8\pi}{3}\delta(r_k)\mathbf{s}_k.\mathbf{I}\right]. \qquad \ldots\ldots(6.5)$$

This is obtained from the Dirac theory of the electron (Fermi 1930). The last term is always zero unless there are s-electrons present, for which case the other terms are zero.

(*e*) The electrostatic interaction between the electrons and the nucleons, giving the nuclear quadrupole interaction, and higher interactions (which are always negligible). The theory of this is not readily accessible, so we give a brief account of it. We let capital letters refer to nucleons and small letters to electrons.

The full coulomb interaction is $\Sigma_{j,N} e_j e_N/(|\mathbf{r}_j - \mathbf{R}_N|)$. This is expanded in powers of $1/r_j$, r_j being the distance of the electron from the nucleus. The first term in the expansion is the term $\Sigma(-Ze^2/r_k)$, which has already been included in (*a*). The second term vanishes because the nucleus does not have an electric dipole moment, and the third term is the one required, being of order $10^{-4}\,\mathrm{cm}^{-1}$ and having the form

$$V_Q = \tfrac{1}{6}\Sigma \frac{e_j e_N}{r_j^5}\left[\begin{array}{l}(r_j^2-3x_j^2)(R_N^2-3X_N^2)+(r_j^2-3y_j^2)(R_N^2-3Y_N^2)\\ \quad+(r_j^2-3z_j^2)(R_N^2-3Z_N^2)\\ \quad+18\{x_jy_jX_NY_N+y_jz_jY_NZ_N+z_jx_jZ_NX_N\}\end{array}\right]. \quad \ldots\ldots(6.6)$$

The parts of this which involve nuclear coordinates can be replaced by equivalent operators, thus $\Sigma_N e_N(R_N^2-3X_N^2)=(-eQ/I(2I-1))\{I(I+1)-3I_x^2\}$, $\Sigma_N e_N X_N Y_N=\{-eQ/I(2I-1)\}\frac{1}{2}\{I_xI_y+I_yI_x\}$ where the multiplicative factor is taken so that Q is defined as usual (Mack 1950). Then with some rearrangement (6.6) becomes:

$$V_Q = \frac{e^2Q}{2I(2I-1)}\Sigma_j\left\{\frac{I(I+1)}{r_j^3}-\frac{3(\mathbf{r}_j.\mathbf{I})^2}{r_j^5}\right\}. \quad \ldots\ldots(6.7)$$

The interaction of the nuclear quadrupole moment with the crystalline electric field is usually neglected.

(*f*) The final term describes the direct interaction of the nuclear moment with the external field: $-\gamma\beta_N\mathbf{H}.\mathbf{I}$. It is of order $10^{-4}\,\mathrm{cm}^{-1}$ and is usually negligible. It can however be used in some cases to determine signs (see § 7.5). A detailed account of these interactions is given by Abragam and Pryce (1951 a).

We now add the term which describes the effect of the crystal field: $\Sigma_j -eV(x_j, y_j, z_j)$ where $V(x, y, z)$ is the potential. In treating its effect by perturbation theory it is first necessary to decide how its order of magnitude compares with the other terms in the Hamiltonian. Three distinct cases have been considered, which we list in order of magnitude, but not in order of importance. First there is the case where some degree of chemical bonding seems to occur between the magnetic ion and its immediate surroundings, typified by the iron group cyanides. Although the theory of this is incomplete it appears that the interaction between the electrons and the surroundings (not necessarily correctly replaced by a crystalline field) is large compared with the coulomb interactions between the electrons, but less than the interaction of the electrons with the nucleus. The second and common case is where the effect of the crystalline field is large compared with the spin–orbit interactions but small compared with the coulomb interactions, typified by the iron group hydrated salts. The last case is where the effect of the field is less than the spin–orbit interactions but greater than the nuclear interactions, typified by rare earth salts.

In describing how the perturbation proceeds it is convenient to deal specifically with one of the above cases and we have chosen to consider the second. The cyanides and rare earths will be reviewed separately (§§ 8 and 9).

Assuming then that the effect of the field is small compared with the coulomb energies, the expression (6.1) is treated by the self-consistent field method (see Condon and Shortley 1951) in which the electrons move independently of each other within the restrictions of the exclusion principle. The eigenstates which this gives are 'configurations', in which the individual electron orbits are specified, e.g. $3d^3$, $3d^24s$, etc. Account is next taken of the correlations between

the electrons (the self-consistent method does not deal fully with the term $\Sigma e^2/r_{ij}$) and at the same time the terms in (6.2) which represent orbit–orbit and spin–spin interactions can be included. The result of this is that the configurations are split into terms, i.e. into levels which are specified by the electron orbits and the total orbital and spin momenta, e.g. $3d^3\ {}^4F$, etc. The next largest perturbation is that which comes from the crystalline field, and the treatment now departs from that of the usual many-electron systems. The problem is then one of finding how the field perturbs the orbital levels, for at this stage no forces which couple spins with orbits have been included. The ground terms of the iron group ions are as given by Hund's rule and have either S, D or F orbital levels. The case when $L=0$ (S state) presents a rather different problem from the others and it is convenient to discuss this separately (§7.3). However before entering on the discussion of how the field splits D and F levels it is necessary to consider the actual field in some detail.

6.2. *The Crystal Field for the Iron Group*

Assuming that the potential satisfies Laplace's equation it can be expanded in spherical harmonics, thus

$$V=\Sigma_{n,\,m}A_n{}^m r^n Y_n{}^m(\theta,\phi) \qquad \ldots\ldots(6.8)$$

where $Y_n{}^m(\theta,\phi)$ is normalized to unity and defined as

$$Y_n{}^m(\theta,\phi)=(-1)^n\left\{\frac{(2n+1)(n-|m|)!}{2(n+|m|)!}\right\}^{1/2}\frac{1}{\sqrt{(2\pi)}}P_n{}^{|m|}(\cos\theta)\,e^{im\phi}$$

where $$P_n{}^m(x)=\frac{(1-x^2)^{m/2}}{2^n n!}\frac{d^{n+m}}{dx^{n+m}}(x^2-1)^n.$$

The number of terms in (6.8) which need to be considered can be considerably reduced as follows. All terms for which $n>4$ will have zero matrix elements and can be omitted. This is because in evaluating an integral such as $\int\chi^* U\psi\,d\tau$ where χ and ψ denote d-electron wave functions, the density $\chi^*\psi$ can also be expanded in spherical harmonics and does not contain harmonics for which $n>4$. If U is a spherical harmonic with $n>4$ the integral vanishes by the orthogonality relations for spherical harmonics. By a similar argument, all terms for which n is odd have zero elements because the product $\chi^*\psi$ is unchanged by the substitution $x, y, z\rightarrow -x, -y, -z$, whereas U reverses sign. The term for $n=0$ can also be dropped because it is an additive constant and is of no theoretical importance. These considerations serve to reduce considerably the number of $A_n{}^m$ which need be retained. To make any further reduction in their number it is necessary to take account of the symmetry properties of the surroundings.

We now consider a model in which the six nearest neighbours are situated near the points $(\pm a, 0, 0)$, $(0, \pm a, 0)$, $(0, 0, \pm a)$, thus forming a distorted octahedron about the magnetic ion. The form which V takes depends on the direction which is chosen for the polar axis. Two directions have been used, and we consider them in turn, beginning with $0z$ as polar axis. If the octahedron is regular the potential at (r, θ, ϕ) will be the same as that at $(r, \theta, \phi+\frac{1}{2}\pi)$, and as $Y_n{}^m(\theta,\phi)$ depends on ϕ according to $\exp(im\phi)$ this means that the only non-vanishing $A_n{}^m$ will be those for which $m=0$ and ± 4 ($|m|\leqslant n\leqslant 4$). Thus V takes the form

$$V=A_2{}^0 r^2 Y_2{}^0(\theta,\phi)+A_4{}^0 r^4 Y_4{}^0(\theta,\phi)+A_4{}^4 r^4\left(Y_4{}^4(\theta,\phi)+A_4{}^{-4} r^4 Y_4{}^{-4}(\theta,\phi)\right.\ldots\ldots(6.9)$$

and, being real, $A_4^4 = \overline{A_4^{-4}}$. This form has been derived on the assumption that the field has fourfold symmetry about $0z$ and would hold if the neighbours did not form a regular octahedron but showed a distortion along $0z$. For a regular octahedron the symmetry is so great that A_2^0 is zero and A_4^0 and A_4^4 are related, so that V becomes

$$V = A_4^0 r^4 \left[Y_4^0(\theta,\phi) + \left(\frac{5}{14}\right)^{1/2} \{Y_4^4(\theta,\phi) + Y_4^{-4}(\theta,\phi)\} \right]. \quad \ldots\ldots(6.10)$$

We now digress to make some remarks about notation. Any surface harmonic $Y_n^m(\theta,\phi)$ for which m is zero shows no dependence on ϕ and is said to have axial symmetry. If $m = \pm 4$ it is said to have tetragonal symmetry, $m = \pm 3$ trigonal symmetry, and $m = \pm 2$ rhombic symmetry. Any form which contains several terms of different symmetry is said to have an overall symmetry which is the greatest symmetry common to all the terms. The special form (6.10) is said to have cubic, or octahedral, symmetry.

Another form for the potentials can be given by changing to Cartesian coordinates, when (6.9) becomes:

$$V = B_2^0(3z^2 - r^2) + B_4^0(35z^4 - 30r^2z^2 + 3r^4) + B_4^4(x^4 - 6x^2y^2 + y^4) \quad \ldots\ldots(6.11)$$

and (6.10) becomes:

$$V = C_4(x^4 + y^4 + z^4 - \tfrac{3}{5}r^4) \quad \ldots\ldots(6.12)$$

where the coefficients are related by numerical factors, thus

$$B_2^0 = \tfrac{1}{4}\left(\frac{5}{\pi}\right)^{1/2} A_2^0, \; B_4^0 = \frac{3}{16\sqrt{\pi}} A_4^0, \; B_4^4 = \tfrac{3}{8}\left(\frac{35}{2\pi}\right)^{1/2} |A_4^4|, \; C_4 = \frac{15}{4\sqrt{\pi}} A_4^0.$$

It is often useful to note that a form having tetragonal symmetry can be expressed as the sum of cubic and axial terms.

We suppose now that the polar axis is taken along the (1, 1, 1) direction. This is a threefold symmetry axis of the octahedron so the general form for V, showing a distortion along the (1, 1, 1) direction, is:

$$V = A_2^0 r^2 Y_2^0(\theta,\phi) + A_4^0 r^4 Y_4^0(\theta,\phi) + A_4^3 r^4 Y_4^3(\theta,\phi) + A_4^{-3} r^4 Y_4^{-3}(\theta,\phi). \quad \ldots(6.13)$$

For a regular octahedron it becomes:

$$V = D_4 r^4 \left[Y_4^0(\theta,\phi) - \left(\frac{10}{7}\right)^{1/2} \{Y_4^3(\theta,\phi) + Y_4^{-3}(\theta,\phi)\} \right]. \quad \ldots\ldots(6.14)$$

This is exactly the same as (6.12), though it looks different because of the different choice of polar axis. A trigonal field is also the sum of cubic and axial terms.

6.3. *Matrix Elements of the Crystal Field*

We now return to the problem of finding the splittings which are caused by the crystal field and will assume that the perturbation consists of a large term of cubic symmetry with smaller terms to describe the departures from cubic symmetry. We begin by taking into account the cubic part of the field acting on either D or F orbital states.

Various ways have been used for treating this problem (see Bethe 1929, Becquerel and Opechowski 1939, Abragam and Pryce 1951 b). We shall describe the method which sets up the secular equation in the usual way for degenerate systems. The polar axis is chosen as axis of quantization with wave functions ψ_m, where m denotes the component of angular momentum in the direction of the axis, and the elements in the secular determinant are the integrals

$$\int \psi_m^* \{\Sigma_j - eV(x_j, y_j, z_j)\} \psi_{m'} \, d\tau.$$

Many of these integrals are zero, as can be seen using the following theorem. If U_N^M is any function which has the same angular variation as $Y_N^M(\theta,\phi)$, the integral $\int\psi_m^* U_N^M \psi_n \, d\tau$ vanishes unless $m=n+M$. The proof of this, which we shall not give, is best obtained using Wigner coefficients (Condon and Shortley 1951). We use the theorem as follows. The cubic field (6.10) is described by terms which have $M=0$ or ± 4 only. The state ψ_0 will therefore have non-zero matrix elements only to states which have $m=0$ or ± 4. But with $L=2$ and 3 there are no states with $m=\pm 4$, so ψ_0 has only a matrix element with itself. This means that it is an eigenstate. Similarly ψ_1 and ψ_{-1} are eigenstates in $L=2$, but in $L=3$ they are coupled to ψ_{-3} and ψ_3 respectively. In both $L=2$ and 3, ψ_2 is coupled to ψ_{-2}. Thus the eigenstates with $L=2$ are ψ_0, ψ_1, ψ_{-1}, (ψ_2, ψ_{-2}) where the notation (ψ_2, ψ_{-2}) means that appropriate linear combinations of ψ_2 and ψ_{-2} must be taken (two mutually orthogonal linear combinations). Similarly for $L=3$ the eigenstates are ψ_0, (ψ_1, ψ_{-3}), (ψ_{-1}, ψ_3), (ψ_2, ψ_{-2}). It will be noticed that in obtaining these results it has been necessary to use only the fact that the potential contains terms for which $M=0, \pm 4$. The tetragonal field also has only these elements so the above considerations apply to it as well. By similar arguments the eigenstates in a trigonal field are

$$L=2;\ \psi_0,\ (\psi_1, \psi_{-2}),\ (\psi_{-1}, \psi_2)$$

$$L=3;\ (\psi_3, \psi_0, \psi_{-3}),\ (\psi_1, \psi_{-2}),\ (\psi_{-1}, \psi_2).$$

In order to obtain the eigenvalues and the coefficients in the linear combinations it is necessary to do more than see which elements are zero—in fact the non-zero elements must be evaluated. This process is greatly simplified by the realization that if $\int\psi_m^* U_N^M \psi_n \, d\tau$ does not vanish, all the other non-vanishing elements for the same N can be found from it by symmetry arguments. The proof of this statement is usually obtained by group theory arguments (Stevens 1952a)—it amounts to showing that any non-zero element $\int\psi_{m'}^* U_N^{M'} \psi_{n'} \, d\tau$ can, by suitable changes of axes, be transformed into a multiple of some known matrix element $\int\psi_m^* U_N^M \psi_n \, d\tau$. The process of replacing operators by multiples of other simpler operators is based on this result. As an example suppose the matrix elements of $U=\Sigma_{\text{electrons}}(35z^4-30r^2z^2+3r^4)$ are required for $3d^2\ {}^3F$. We begin by calculating $\int\psi_3^* U\psi_3 \, d\tau$. This presents little difficulty because the product of single electron states $\{\overset{+}{2}\overset{+}{1}\}$ (using a symbolic representation in which the curly bracket means that the state is antisymmetric) has $L_z=3$ and $S_z=1$ and can be taken as ψ_3. Thus the original integral becomes the sum of two single d-electron integrals

$$\int\psi_3^* U\psi_3 \, d\tau = \int\{\overset{+}{2}\overset{+}{1}\}^* U\{\overset{+}{2}\overset{+}{1}\} \, d\tau$$
$$= \int\chi_2^*(35z^4-30r^2z^2+3r^4)\chi_2 \, d\tau + \int\chi_1^*(35z^4-30r^2z^2+3r^4)\chi_1 \, d\tau.$$

The wave functions of d-electrons have the forms $\chi_m=f(r)Y_2^m(\theta,\phi)$ where $f(r)$ denotes the radial wave function and m denotes the component of angular momentum in the direction of the axis. The angular parts of these integrals are readily evaluated, but as $f(r)$ is usually unknown the radial integral cannot be performed, and instead the following definition is used:

$$\overline{r^n} = \int[f(r)]^2 r^n r^2 \, dr. \qquad \ldots\ldots(6.15)$$

It follows that

$$\int\psi_3^* U\psi_3 \, d\tau = -8\overline{r^4}/7. \qquad \ldots\ldots(6.16)$$

The other elements can be obtained in a similar way, by expressing each state of the system in terms of product states. This is a tedious process and can, in fact, be largely avoided, since it can be shown that the matrix elements of $\Sigma(35z^4-30r^2z^2+3r^4)$ are proportional to the matrix elements of a similar operator in which x, y and z are everywhere replaced by L_x, L_y and L_z (Stevens 1952a). The only difficulty in making this replacement is that whereas x, y and z commute, L_x, L_y and L_z do not—account is taken of this by replacing Σxy, for example, by $\frac{1}{2}(L_xL_y+L_yL_x)$, and similarly for more complicated expressions. Thus the elements of $\Sigma(35z^4-30r^2z^2+3r^4)$ are equal to the elements of $\overline{r^4}\beta[35L_z^4-30L(L+1)L_z^2+25L_z^2-6L(L+1)+3L^2(L+1)^2]$ where β is the proportionality factor. Knowing the value of one element (6.16), β is readily found to equal $-2/315$. Suppose now that the matrix elements of $\Sigma(x^4-6x^2y^2+y^4)$ are required. It is not necessary to evaluate any more integrals because its elements are the same as those of $\overline{r^4}\beta[\frac{1}{2}(L_x+iL_y)^4+\frac{1}{2}(L_x-iL_y)^4]$ since $x^4-6x^2y^2+y^4=\frac{1}{2}(x+iy)^4+\frac{1}{2}(x-iy)^4$. In view of the power of this method of finding the matrix elements of the crystal field we give, in table 1, a number of equivalent operators and the factors for the ground terms of the iron group ions. Using these it is easy to set up the secular equation for any trigonal or tetragonal crystal field perturbation. In the case of cubic symmetry the energy levels are as given in fig. 6 and the states are given below for the two choices of direction for the axis of quantization. We use Δ for the overall splitting, so that $\Delta=6\beta$ for $L=2$ and 54β for $L=3$.

$L=2$

Energy	Tetragonal	Trigonal
$\frac{18}{5}\beta$	ψ_0	$\sqrt{\frac{1}{3}}\psi_{-2}-\sqrt{\frac{2}{3}}\psi_1$
	$\{\psi_2+\psi_{-2}\}/\sqrt{2}$	$\sqrt{\frac{1}{3}}\psi_2+\sqrt{\frac{2}{3}}\psi_{-1}$
$-\frac{12}{5}\beta$	ψ_1	ψ_0
	ψ_{-1}	$\sqrt{\frac{2}{3}}\psi_{-2}+\sqrt{\frac{1}{3}}\psi_1$
	$\{\psi_2-\psi_{-2}\}/\sqrt{2}$	$\sqrt{\frac{2}{3}}\psi_2-\sqrt{\frac{1}{3}}\psi_{-1}$

$L=3$

Energy	Tetragonal	Trigonal
18β	$\sqrt{\frac{3}{8}}\psi_{-1}+\sqrt{\frac{5}{8}}\psi_3$	$\frac{2}{3}\psi_0-\frac{1}{3}\sqrt{\frac{5}{2}}(\psi_3-\psi_{-3})$
	$\sqrt{\frac{3}{8}}\psi_1+\sqrt{\frac{5}{8}}\psi_{-3}$	$\sqrt{\frac{5}{6}}\psi_2+\sqrt{\frac{1}{6}}\psi_{-1}$
	ψ_0	$\sqrt{\frac{5}{6}}\psi_{-2}-\sqrt{\frac{1}{6}}\psi_1$
-6β	$\sqrt{\frac{5}{8}}\psi_{-1}-\sqrt{\frac{3}{8}}\psi_3$	$(\psi_3+\psi_{-3})/\sqrt{2}$
	$\sqrt{\frac{5}{8}}\psi_1-\sqrt{\frac{3}{8}}\psi_{-3}$	$\sqrt{\frac{1}{6}}\psi_2-\sqrt{\frac{5}{6}}\psi_{-1}$
	$(\psi_2+\psi_{-2})/\sqrt{2}$	$\sqrt{\frac{1}{6}}\psi_{-2}+\sqrt{\frac{5}{6}}\psi_1$
-36β	$(\psi_2-\psi_{-2})/\sqrt{2}$	$\frac{1}{3}\sqrt{2}(\psi_2-\psi_{-2})-\frac{1}{3}\sqrt{5}\psi_0$

§7. THE SPIN-HAMILTONIAN

7.1. *Derivation*

Having dealt with the effect of the crystal field on the orbital levels we now pass to the consideration of the other perturbations. The first of these is the term $\Sigma a_{jk}\mathbf{l}_j.\mathbf{s}_k$ occurring in (6.2). It has the property that it couples spins with orbits, so that the spin degeneracy can no longer be neglected. Further,

Table 1

$$L_+ = L_x + iL_y, \qquad L_- = L_x - iL_y$$

$$\Sigma(3z^2 - r^2) = \alpha\overline{r^2}\{3L_z^2 - L(L+1)\}$$

$$\Sigma(x^4 + y^4 + z^4 - \tfrac{3}{5}r^4) = \beta\overline{r^4}\{L_x^4 + L_y^4 + L_z^4 - \tfrac{1}{5}L(L+1)(3L^2 + 3L - 1)\}$$

$$= \frac{\beta\overline{r^4}}{20}\{35L_z^4 - 30L(L+1)L_z^2 + 25L_z^2 - 6L(L+1) + 3L^2(L+1)^2\} + \frac{\beta\overline{r^4}}{5}\{L_+^4 + L_-^4\}$$

$$\Sigma(x^4 - 6x^2y^2 + y^4) = \tfrac{1}{2}\beta\overline{r^4}\{L_+^4 + L_-^4\}$$

$$\Sigma(35z^4 - 30r^2z^2 + 3r^4) = \beta\overline{r^4}\{35L_z^4 - 30L(L+1)L_z^2 + 25L_z^2 - 6L(L+1) + 3L^2(L+1)^2\}$$

$$\Sigma(x^2y - xy^2) = -\tfrac{1}{4}i\beta\overline{r^4}\{L_+^4 - L_-^4\}$$

$$\Sigma x(x^2 - 3xy^2) = \tfrac{1}{8}\beta\overline{r^4}\{L_z(L_+^3 + L_-^3) + (L_+^3 + L_-^3)L_z\}$$

$$\Sigma x(y^2 - 3x^2y) = \tfrac{1}{4}i\beta\overline{r^4}\{L_z(L_+^3 - L_-^3) + (L_+^3 - L_-^3)L_z\}$$

$$\alpha = -\frac{2}{21}, \quad \beta = \frac{2}{63} \quad \text{for} \begin{cases} 3d^1\ {}^2D \\ 3d^6\ {}^5D \end{cases} \qquad \alpha = \frac{2}{21}, \quad \beta = -\frac{2}{63} \quad \text{for} \begin{cases} 3d^4\ {}^5D \\ 3d^9\ {}^2D \end{cases}$$

$$\alpha = -\frac{2}{105}, \quad \beta = -\frac{2}{315} \quad \text{for} \begin{cases} 3d^2\ {}^3F \\ 3d^7\ {}^4F \end{cases} \qquad \alpha = \frac{2}{105}, \quad \beta = \frac{2}{315} \quad \text{for} \begin{cases} 3d^3\ {}^4F \\ 3d^8\ {}^3F \end{cases}$$

In general: $$\alpha = \mp\frac{2(2l+1-4S)}{(2l-1)(2l+3)(2L-1)}$$

$$\beta = \mp\frac{3(2l+1-4S)\{-7(l-2S)(l-2S+1)+3(l-1)(l+2)\}}{(2l-3)(2l-1)(2l+3)(2l+5)(L-1)(2L-1)(2L-3)}$$

where $-$ sign is used in the first half of the shell and $+$ sign is used in the second half of the shell.

provided that we use only states which are eigenstates of **L** and **S** (that is, restrict our consideration to states coming from the ground term), this perturbation is equivalent to the usual spin–orbit interaction given by $\lambda\mathbf{L}.\mathbf{S}$ (Condon and Shortley 1951). Similarly the spin–spin interaction (6.3) takes the form

$$-\rho[(\mathbf{L}.\mathbf{S})^2 + \tfrac{1}{2}(\mathbf{L}.\mathbf{S}) - \tfrac{1}{3}L(L+1)S(S+1)], \qquad \ldots\ldots(7.1)$$

the magnetic interaction (6.5) becomes

$$\left(\overline{\frac{1}{r^3}}\right)2\gamma\beta\beta_N[\mathbf{L} + \xi\{L(L+1)(\mathbf{S}.\mathbf{I}) - \tfrac{3}{2}(\mathbf{L}.\mathbf{S})(\mathbf{L}.\mathbf{I}) - \tfrac{3}{2}(\mathbf{L}.\mathbf{I})(\mathbf{L}.\mathbf{S})\} - \kappa\mathbf{S}.\mathbf{I}] \quad \ldots(7.2)$$

and the quadrupole interaction (6.7) becomes

$$\left(\overline{\frac{1}{r^3}}\right)\eta[3(\mathbf{L}.\mathbf{I})^2 + \tfrac{3}{2}(\mathbf{L}.\mathbf{I}) - L(L+1)I(I+1)] \qquad \ldots\ldots(7.3)$$

where $$\xi = \frac{2l+1-4S}{S(2l-1)(2l+3)(2L-1)}, \qquad \eta = \pm 2S\xi \qquad (l=2)$$

and the sign of η depends on whether the shell is less than or more than half full. These expressions are quoted without proof by Abragam and Pryce (1951 a). The proofs are very similar and by way of example we shall indicate how the term

$$\sum_{\substack{k \\ \text{electrons}}}\left\{\frac{3(\mathbf{r}_k.\mathbf{s}_k)(\mathbf{r}_k.\mathbf{I})}{r_k^5} - \frac{(\mathbf{s}_k.\mathbf{I})}{r_k^3}\right\} \qquad \ldots\ldots(7.4)$$

which occurs in (6.5) is reduced to a form

$$\left(\overline{\frac{1}{r^3}}\right)\xi\{L(L+1)(\mathbf{S}.\mathbf{I}) - \tfrac{3}{2}(\mathbf{L}.\mathbf{S})(\mathbf{L}.\mathbf{I}) - \tfrac{3}{2}(\mathbf{L}.\mathbf{I})(\mathbf{L}.\mathbf{S})\}. \qquad \ldots\ldots(7.5)$$

The proof depends essentially on replacing the variables by appropriate angular momentum operators in much the same way as is done for the crystal field.

Consider the term in I_x. This is multiplied by

$$\Sigma\left[s_x\left(-\frac{1}{r^3}+\frac{3x^2}{r^5}\right)+\frac{3zx}{r^5}s_z+\frac{3xy}{r^5}s_y\right]_k. \qquad \ldots\ldots(7.6)$$

The parts which involve spin variables transform under rotation like the components of a vector, and therefore within a manifold in which S is a constant (7.6) can be replaced by

$$\tau\left[S_x\Sigma\left(\frac{3x^2-r^2}{r^5}\right)_k+S_z\Sigma\left(\frac{zx}{r^5}\right)_k+S_y\Sigma\left(\frac{xy}{r^5}\right)_k\right]$$

where τ is a factor which has yet to be determined. Similarly the parts in (7.6) which involve position coordinates transform like the second order spherical harmonics. Within a manifold in which $\mathbf{L}$ is constant they can be replaced by

$$\alpha\tau\left(\overline{\frac{1}{r^3}}\right)[S_x\{3L_x^2-L(L+1)\}+\tfrac{1}{2}S_x(L_xL_z+L_zL_x)+\tfrac{1}{2}S_y(L_xL_y+L_yL_x)]$$

where α is the same factor as is used in replacing the second order crystal field. With some rearrangement this expression together with the terms involving I_x and I_y can be written in the form (7.5). The factors α and τ are found by taking special cases and their product, which equals $-\xi$, is given by the above formula.

The final result of the various transformations of the different perturbations is to produce a total perturbation

$$\begin{aligned}V+(\lambda-\tfrac{1}{2}\rho)(\mathbf{L}.\mathbf{S})-\rho(\mathbf{L}.\mathbf{S})^2+\beta\mathbf{H}.(\mathbf{L}+2\mathbf{S})\\+p[(\mathbf{L}.\mathbf{I})+\{\xi L(L+1)-\kappa\}(\mathbf{S}.\mathbf{I})-\tfrac{3}{2}\xi(\mathbf{L}.\mathbf{S})(\mathbf{L}.\mathbf{I})-\tfrac{3}{2}\xi(\mathbf{L}.\mathbf{I})(\mathbf{L}.\mathbf{S})]\\+q'\{(\mathbf{L}.\mathbf{I})^2+\tfrac{1}{2}(\mathbf{L}.\mathbf{I})\}-\gamma\beta_N(\mathbf{H}.\mathbf{I}) \qquad \ldots\ldots(7.7)\end{aligned}$$

where V denotes any crystalline field which has not already been considered.

A very useful method of carrying out the perturbation calculation has been given by Pryce (1950 b) (for an earlier example see Van Vleck 1940), which we now discuss. The method is only applied when the lowest orbital state is a singlet, and consists of carrying out a formal perturbation treatment in which the operators referring to spin and nuclear variables are treated as non-commuting algebraic quantities. That is, no representation is chosen for either $\mathbf{S}$ or $\mathbf{I}$. The result of this is that instead of obtaining the perturbed energy as usual, an expression is obtained which involves the components of $\mathbf{S}$ and $\mathbf{I}$. This is the so-called spin-Hamiltonian and the actual energies are the eigenvalues of this new operator.

In describing how the method is used the form (7.7) is unnecessarily complicated, and we shall instead take a simpler perturbation:

$$\mathscr{H}'=\beta\mathbf{H}.(\mathbf{L}+2\mathbf{S})+\lambda\mathbf{L}.\mathbf{S}+A\mathbf{I}.\mathbf{S}. \qquad \ldots\ldots(7.8)$$

It is also convenient to use the Dirac notation and to take the ground orbital state as $|0\rangle$ and the excited states as $|n\rangle$, with energies $E(n)$, where $n=1, 2, 3$, etc. (It may be remarked that the Dirac notation is such that an expression $\langle n|X|m\rangle$ is equal to a matrix element which in the Schrödinger representation would be written $\int\psi_n^*X\psi_m\,d\tau$.) Applying first order perturbation theory there is no difficulty about degeneracy as it is assumed that the ground state (orbit only) is non-degenerate. Thus $\langle 0|\mathscr{H}'|0\rangle=\langle 0|\beta\mathbf{H}.(\mathbf{L}+2\mathbf{S})+\lambda\mathbf{L}.\mathbf{S}+A\mathbf{I}.\mathbf{S}|0\rangle$. In the terms $\langle 0|\beta\mathbf{H}.2\mathbf{S}|0\rangle$ and $\langle 0|A\mathbf{I}.\mathbf{S}|0\rangle$ there are no operators acting on the orbits and $\langle 0|0\rangle=1$, so they give

$$2\beta\mathbf{H}.\mathbf{S}+A\mathbf{I}.\mathbf{S}. \qquad \ldots\ldots(7.9)$$

The term $\langle 0|\beta\mathbf{H}.\mathbf{L}+\lambda\mathbf{L}.\mathbf{S}|0\rangle$ similarly equals

$$(\beta H_x+\lambda S_x)\langle 0|L_x|0\rangle+(\beta H_y+\lambda S_y)\langle 0|L_y|0\rangle+(\beta H_z+\lambda S_z)\langle 0|L_z|0\rangle$$

and vanishes because $\langle 0|L_x|0\rangle=\langle 0|L_y|0\rangle=\langle 0|L_z|0\rangle=0$. This is a consequence of a general theorem that a singlet orbital state has no orbital moment. We now proceed to the next order in perturbation, and consider

$$-\sum_{n\neq 0}\frac{|\langle 0|\beta\mathbf{H}.(\mathbf{L}+2\mathbf{S})+\lambda\mathbf{L}.\mathbf{S}+A\mathbf{I}.\mathbf{S}|n\rangle|^2}{E(n)-E(0)}.$$

The terms which do not involve orbital variables vanish because $\langle 0|n\rangle$ is zero when n is non-zero—in particular there are no terms from this form which involve nuclear variables. The non-vanishing elements such as $\langle 0|L_x|n\rangle$, $\langle 0|L_y|n\rangle$, $\langle 0|L_z|n\rangle$ must be evaluated in any particular case, but in the present connection it is sufficient to note that the final result will be quadratic in the components of $\mathbf{H}$ and $\mathbf{S}$, and will therefore take the form

$$-\beta^2\Lambda_{ij}H_iH_j-\beta f_{ij}H_iS_j+d_{ij}S_iS_j \qquad \ldots\ldots(7.10)$$

where $i, j=x, y, z$ and Λ_{ij}, f_{ij} and d_{ij} are factors. It is not usually necessary to go to any higher order in perturbation theory so, collecting together the two contributions (7.9) and (7.10), we have the result that the properties of the spin and nuclear degeneracies of the lowest singlet orbital are described by the properties of the 'spin-Hamiltonian'

$$\mathscr{H}_s=\beta(2\delta_{ij}-f_{ij})H_iS_j+A\mathbf{I}.\mathbf{S}-\beta^2\Lambda_{ij}H_iH_j+d_{ij}S_iS_j. \qquad \ldots\ldots(7.11)$$

This can also be written in the shorter form

$$\mathscr{H}_s=\beta\mathbf{H}.g.\mathbf{S}+A\mathbf{I}.\mathbf{S}-\beta^2\mathbf{H}.\Lambda.\mathbf{H}+\mathbf{S}.D.\mathbf{S} \qquad \ldots\ldots(7.12)$$

where $g=2\delta-f$, Λ and D are tensors. This spin-Hamiltonian has been derived for a fairly simple perturbation. If the general form (7.7) is used, with the simplifying assumption that the spin–spin and electron–nucleus interactions need only be taken to first order, the general spin-Hamiltonian becomes

$$\mathscr{H}_s=d_{ij}S_iS_j+\beta g_{ij}H_iS_j+A_{ij}S_iI_j+p_{ij}I_iI_j-\gamma\beta_N\mathbf{H}.\mathbf{I}-\beta^2\Lambda_{ij}H_iH_j. \qquad \ldots\ldots(7.13)$$

This can also be written in the form

$$\mathscr{H}_s=\mathbf{S}.D.\mathbf{S}+\beta\mathbf{H}.g.\mathbf{S}+\mathbf{S}.T.\mathbf{I}+\mathbf{I}.P.\mathbf{I}-\gamma\beta_N\mathbf{H}.\mathbf{I}-\beta^2\mathbf{H}.\Lambda.\mathbf{H}. \qquad \ldots(7.14)$$

7.2. *General Discussion*

In the preceding section a general derivation of the spin-Hamiltonian was given. For a general discussion it is convenient to assume that the symmetry of the crystalline field is either tetragonal or trigonal. It is clear that the spin-Hamiltonian must show the same symmetry as the field and for these cases it takes the form

$$\mathscr{H}_s=DS_z^2+\beta[g_\parallel H_zS_z+g_\perp(H_xS_x+H_yS_y)]+AI_zS_z+B(I_xS_x+I_yS_y)+PI_z^2-\gamma\beta_N\mathbf{H}.\mathbf{I}-\beta^2\mathbf{H}.\Lambda.\mathbf{H}. \qquad \ldots\ldots(7.15)$$

The various terms can be interpreted as follows. The terms in $\mathbf{H}$ and $\mathbf{S}$ show that the spin is no longer completely free, but can either be regarded as free and precessing about the field $\{g_\perp H_x, g_\perp H_y, g_\parallel H_z\}$, or as having an anisotropic magnetic moment $\{g_\perp S_x, g_\perp S_y, g_\parallel S_z\}$. The term DS_z^2 describes how the levels behave in zero magnetic field in the absence of the nuclear interaction and comes from the second order effects of the crystal field and the spin–orbit coupling, and from the spin–spin interaction. The terms in $\mathbf{S}$ and $\mathbf{I}$ come from the

magnetic interactions between the electrons and the nucleus and are anisotropic in the same way as the electronic moment is anisotropic. The term PI_z^2 comes from the quadrupole interaction and shows that this is also anisotropic. The direct interaction between the magnetic field and the nuclear moment takes its usual form $-\gamma\beta_N \mathbf{H}.\mathbf{I}$, and the final term $-\beta^2 \mathbf{H}.\Lambda.\mathbf{H}$ does not depend either on **S** or **I** and can be neglected as an additive constant. It needs to be considered in evaluating magnetic susceptibilities as it gives rise to temperature independent paramagnetism. In concluding this explanation of the terms we would emphasize that the form (7.14) of the spin-Hamiltonian has been deduced on the assumption that the lowest orbital level is a singlet.

It may be noted that the Hamiltonian (7.15) has complete axial symmetry, and hence so also has the resonance spectrum, showing that the exact nature of the crystalline field symmetry (e.g. trigonal or tetragonal) cannot be deduced from the experiments. The latter is only possible if powers of S_z, etc. greater than 2 occur, as happens for ions in S-states (see §7.3).

The importance of the spin-Hamiltonian lies in the fact that its employment makes it possible to give a description of the resonance properties of an ion in terms of a relatively small number of constants. Without this it is necessary to give complicated formulae and diagrams of how energy levels diverge as H is varied, as was done in many of the earlier papers on the subject. This does not imply that everything is now very simple, for it is often a complicated process to find the eigenvalues of a given spin-Hamiltonian, but it does imply that all the relevant information can be given in a compact form. The realization of this has served to bring about a fairly sharp division between theoretical and experimental work—experiment is aimed at finding the spin-Hamiltonian, and theory at passing from a model to a spin-Hamiltonian, or in the reverse direction. This division is further emphasized by the knowledge that it is often possible to guess the form for the spin-Hamiltonian with practically no detailed knowledge of the crystal field. We proceed to the discussion of this.

Suppose we ask what the low-lying energy levels will look like if we remove the restriction that the ground orbital level is non-degenerate. We first note that the splittings given by the interactions with the nucleus are very small ($\sim 10^{-2}\,\mathrm{cm}^{-1}$) and as a first approximation can be neglected. Thus each electronic state has $(2I+1)$-fold degeneracy coming from the nuclear spin. We next consider the conclusions to which the following two theorems point. First there is Kramers' theorem (Kramers 1930) which states that if a system containing an odd number of electrons is placed in an external electric field every energy level will have even degeneracy. Secondly, there is the theorem of Jahn and Teller (1937) which states that, apart from linear complexes, a complex which has a degenerate ground state will spontaneously distort in such a way that the degeneracy is removed. Physically, this is because by so doing the system can reduce its energy. Of course if there is only Kramers' degeneracy there is no distortion which will remove it. The effect of these two theorems is that for the most complicated systems it is reasonable to suppose that the ground electronic state (orbit and spin) will be a singlet when the number of electrons is even and a doublet when the number is odd. When there is an orbital singlet lowest the splitting of the spin degeneracy is usually fairly small and the treatment given above shows that the spin-Hamiltonian should be that appropriate to a spin equal to the actual spin. When, however, there is an orbitally degenerate

level lowest the departures of the field from cubic symmetry and the spin–orbit coupling taken together usually produce splittings of order 100 cm^{-1}, each level having the least possible degeneracy. With an even number of electrons the electronic state will be a singlet and no resonance will be observed. With an odd number of electrons the ground electronic state will have the two-fold Kramers' degeneracy and there is the possibility that resonance will be obtained. The next excited state will, in general, be some 100 cm^{-1} higher and the ground level can be regarded as an isolated doublet, and described in terms of an effective spin $\frac{1}{2}$. In a magnetic field this behaves like $\beta\mathbf{H}.g.\mathbf{S}(S=\frac{1}{2})$, and when allowance is made for the nuclear interactions the usual spin-Hamiltonian is obtained. In particular the form (7.15) is obtained for either tetragonal or trigonal symmetry. The only cases in the iron group where this method of deriving the spin-Hamiltonian has failed is for the case of ions in S states, and these we discuss separately (§7.3).

To summarize, when the ground orbital state is a singlet in the cubic field the spin-Hamiltonian appropriate to the actual spin must be used. When the ground orbital state is degenerate, an effective spin $\frac{1}{2}$ is used when the number of electrons is odd and no resonances are expected when the number of electrons is even.

Before leaving this general consideration of spin-Hamiltonians it may be useful to point out that if spin–spin interaction is neglected there is a relation between the terms which give the zero field splittings and the anisotropy in the g's. The electronic terms in the Hamiltonian can be written $2\beta\mathbf{H}.\mathbf{S}+\mathbf{L}.(\beta\mathbf{H}+\lambda\mathbf{S})$, of which the first term goes through unaltered in the first order spin-Hamiltonian whereas the second is zero. In second order, however, the first term gives zero and the second gives rise to anisotropic terms; in particular the components of $\beta\mathbf{H}$ will always be accompanied by the components of $\lambda\mathbf{S}$. Suppose for example that the term $\beta\mathbf{H}.g.\mathbf{S}$ is known to occur in a spin-Hamiltonian. Writing it as $2\beta\mathbf{H}.\mathbf{S}+\beta\mathbf{H}.(g-2).\mathbf{S}$, the above considerations imply that the form

$$2\beta\mathbf{H}.\mathbf{S}+(2\lambda)^{-1}(\beta\mathbf{H}+\lambda\mathbf{S}).(g-2).(\beta\mathbf{H}+\lambda\mathbf{S})$$

is to be expected, which gives both the zero field splittings and the temperature independent paramagnetic terms.

7.3. *S-State Ions*

So far we have not considered what happens when the orbital state of the free ion has no degeneracy. Assuming that the Hund rule holds this will happen whenever the unfilled shell is half filled, giving in the iron group $3d^5$ ^6S and in the rare earths $4f^7$ ^8S. It is convenient to discuss a particular case, so we take $3d^5$ ^6S. If we proceed as in the other iron group ions and apply the crystal field on the orbital state there can of course be no splitting, since there is no orbital degeneracy. In fact the only way in which a splitting can be obtained is to include a coupling between spins and orbits and regard the problem as one in which the ground level has sixfold degeneracy. It is found that using direct spin–orbit interactions and the crystal field (Van Vleck and Penney 1934) it is necessary to go to a high order in perturbation theory before any splitting of the levels is obtained. Resonance has shown that small splittings do exist, but Abragam and Pryce (1951 a) came to the conclusion that these splittings are in fact too big to be explained by the above process. They have suggested

instead that they are due to a spin–spin interaction of the form (6.3), which also couples orbits with spins, since it depends on position variables. Physically their idea was as follows. In the crystal field, even though the ion is in an S state, there will be some distortion of the orbits. Suppose therefore that instead of the charge cloud being a perfect sphere it is slightly ellipsoidal. Then the dipole–dipole energy of the spins varies with their orientation with respect to the axes of the ellipsoid and thus the eigenvalues depend on the spin orientation. They proceeded to work out the magnitude of the splitting on this mechanism and were able to put forward convincing arguments that this is indeed the correct explanation. In the case of gadolinium ethyl sulphate (see § 9) it seems likely that the same process is again responsible for the splittings.

The Spin-Hamiltonian for S States. We now consider the problem of finding the spin-Hamiltonian for S-state ions. The previous considerations do not help very much because this is an exceptional case and the best we can do is to use a general form for it. We begin by considering only the electronic splittings, and neglect the nuclear interactions. With $S=\frac{1}{2}$, all terms like S_z^2, S_x^4 reduce to constants, and in fact any expression in S_x, S_y and S_z can be reduced either to a constant or a linear expression in S_x, S_y and S_z. Similarly if $S=1$ all formulae can be reduced to expressions in S_x, S_y, S_z, S_z^2, $(S_x+iS_y)^2$, $(S_x-iS_y)^2$, S_zS_x and S_yS_x (Stevens 1952b). This process can be continued for higher spin values and for $3d^5\ {}^6S$ we want to know what expressions are needed to span a general expression in S_x, S_y and S_z when $S=5/2$. Actually we do not need to know the full answer to this because in the absence of a magnetic field, all theoretical considerations suggest that the spin-Hamiltonian should contain only even powers of S_x, S_y and S_z, and also the spin-Hamiltonian must reflect the overall symmetry of the problem. Suppose then that the crystal field is of cubic symmetry. If the spin-Hamiltonian contains S_x^2 it must contain similar amounts of S_y^2 and S_z^2, and as $S_x^2+S_y^2+S_z^2=S(S+1)$ this is just a constant and can be omitted. Similarly S_x^4 implies $S_x^4+S_y^4+S_z^4$ and $S_x^2S_y^2$ implies $S_x^2S_y^2+S_y^2S_z^2+S_z^2S_x^2+S_y^2S_x^2+S_z^2S_y^2+S_x^2S_z^2$. This latter form does not need to be used however as the relation

$$(S_x^2+S_y^2+S_z^2)^2=S_x^4+S_y^4+S_z^4 \\ +(S_x^2S_y^2+S_y^2S_z^2+S_z^2S_x^2+S_y^2S_x^2+S_z^2S_y^2+S_x^2S_z^2)$$

reduces it to the earlier form. Passing now to terms in sixth and higher powers of S_x, S_y and S_z it can be shown that any term can be reduced to one of those of lesser degree which have already been considered. (This is the extension of our discussion of $S=\frac{1}{2}$ and 1 and can be proved by the theorems of § 6.3.) Thus with cubic symmetry and no magnetic field, for spin 5/2, the spin-Hamiltonian is

$$\tfrac{1}{6}a[S_x^4+S_y^4+S_z^4-\tfrac{1}{5}S(S+1)(3S^2+3S-1)].$$

Here we have used $\frac{1}{6}a$ so that a has its usual meaning and have added constant terms to make the expression analogous with the corresponding equivalent operator in table 1. The complex interactions which produce the electronic splittings will have very little effect on the magnetic properties and the nuclear interactions, so the complete spin-Hamiltonian will be

$$g\beta\mathbf{H}.\mathbf{S}+\tfrac{1}{6}a[S_x^4+S_y^4+S_z^4-\tfrac{1}{5}S(S+1)(3S^2+3S-1)]+A\mathbf{I}.\mathbf{S}-\gamma\beta_N\mathbf{H}.\mathbf{I}.$$

If instead there is only tetragonal or trigonal symmetry in the crystal field the general spin-Hamiltonian is

$$\beta[g_\parallel H_1 S_1 + g_\perp (H_2 S_2 + H_3 S_3)] + \tfrac{1}{6}a[S_x^4 + S_y^4 + S_z^4 - \tfrac{1}{5}S(S+1)(3S^2+3S-1)]$$
$$+ b[35S_1^4 - 30S(S+1)S_1^2 + 25S_1^2 - 6S(S+1) + 3S^2(S+1)^2]$$
$$+ D[S_1^2 - \tfrac{1}{3}S(S+1)] + AS_1 I_1 + B(S_2 I_2 + S_3 I_3)$$
$$+ P[I_1^2 - \tfrac{1}{3}I(I+1)] - \gamma\beta_N \mathbf{H}.\mathbf{I}$$

where 01, 02 and 03 are orthogonal axes with 01 in the direction of the distortion. The extra spin operators which now appear are really of the form $\alpha S_1^2 + \beta S_1^4$, but we have written them as above to conform with the equivalent operators given in table 1. This is of some convenience if it becomes necessary to change axes of coordinates. Since the electronic splittings and the other effects are practically unrelated, $g_\parallel \simeq g_\perp$ and $A \simeq B$. In addition, if there are no unpaired s-electron states admixed into the d^5 states, A is very small.

For the case of gadolinium it is necessary to use terms up to the sixth degree in S_x, S_y and S_z. The spin-Hamiltonian is given in the discussion of the gadolinium ion in §9.2 and will not be further discussed here.

7.4. *The Anomalous Hyperfine Structure*

It is implicit in our discussion of the spin-Hamiltonian that, if the crystal field is known, it should be possible to calculate the ratio $A : B$, since by taking the ratio γ and $\overline{(1/r^3)}$ are eliminated. When this was tried for the iron group ions, assuming that the electronic state was of the form $3d^n$, no agreement could be found. In particular no hyperfine structure was expected for manganese ($3d^5\ ^6S$), and in fact a large isotropic hyperfine structure was found. Abragam and Pryce (1951 a) suggested that the discrepancy arises because the ground states of the ions are not built entirely from configurations of type $3d^n$, but contain admixtures of type $3s\ 3d^n\ 4s$. These are presumed to be introduced by configurational interactions, i.e. $\Sigma e^2/r_{ij}$. States of this type are particularly important because when there are unpaired s-electrons present the second term in (6.5) must be used and this is large for 3s electrons. Thus small admixtures of this type can give anomalously large contributions to the hyperfine structure and make no noticeable difference in any of the other properties. By introducing this hypothesis it was possible to account for the observed values of A and B as well as the ratio $A:B$, and although the correctness of this explanation has sometimes been questioned no other detailed explanation has been given. In this connection it may be remarked that the Abragam and Pryce effect would also be present in the free ion hyperfine structure, whereas some of the other explanations (e.g. that the ion picks up an extra electron from its surroundings and this goes into the 4s shell) would not then be present.

Experimentally it turns out that the contribution to the magnetic field at the nucleus from unpaired s-electron configurations is remarkably constant throughout the iron group. This is illustrated by the following values of $-2\beta\beta_N\kappa\overline{(1/r^3)}$ (κ is the coefficient of the isotropic term $\mathbf{S}.\mathbf{I}$ in eqn. (7.2)) for hydrated salts such as double sulphates, where the waters of crystallization are next to the paramagnetic ions: Cu^{2+} 0·006 cm^{-1}, Co^{2+} 0·0054 cm^{-1}, Mn^{2+} 0·0067 cm^{-1}, V^{2+} 0·0060 cm^{-1}. Where the ion is more closely bound, as in the cyanides (or Mn^{2+} in ZnS), the contribution seems to be about 30% lower, but there is less experimental evidence.

In the case of the rare earth ions it has been found possible to predict the ratio $A:B$ without having to postulate admixtures of unpaired s-electrons.

7.5. *The Resonance Spectrum*

From the spin-Hamiltonian it is possible to evaluate the energy levels and the allowed transitions. General formulae can be given for these only in restricted cases. Bleaney (1951 b) has given expressions for the allowed transitions in strong fields (that is, when the term $\beta\mathbf{H}.g.\mathbf{S}$ is much greater than any other term in the Hamiltonian) for the Hamiltonian (7.15), which is limited to the case of axial symmetry, and omits terms of higher powers than S^2. His formulae will not be repeated here, but we will point out one or two of the main features.

The spectrum in sufficiently strong fields will be centred on the point given by the equation $h\nu = g\beta H$. If g is anisotropic, and H is applied at an angle θ to the symmetry axis

$$g^2 = g_\parallel^2 \cos^2\theta + g_\perp^2 \sin^2\theta. \qquad \ldots\ldots(7.16)$$

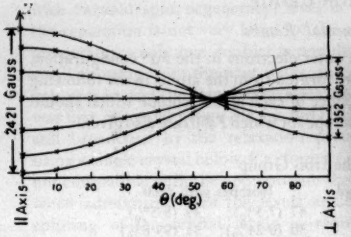


Fig. 7. Plot showing positions of the seven fine structure lines of gadolinium ethyl sulphate as a function of the angle θ between the magnetic field and the axis of the crystalline field. The variation is predominantly as $3\cos^2\theta - 1$ (*after* H. E. D. Scovil).

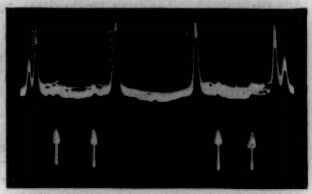


Fig. 8. Oscillogram of hyperfine structure of copper ($S=1/2$, $I=3/2$). The strong lines due to the two isotopes are clearly resolved at the ends of the spectrum. The arrows indicate weak transitions arising from an electric quadrupole interaction (*after* Bleaney, Bowers and Ingram 1951).

If a fine structure is present due to the term DS_z^2, then it will consist of $2S$ equally spaced lines, whose intensity is greatest for the central lines, and smallest for the outermost lines. The separation varies as $3\cos^2\theta - 1$, as is shown in fig. 7 for gadolinium. Here higher terms than S_z^2 are present, but they are small and cause the lines to deviate only slightly from equal spacing. When hyperfine structure is present, due to a nucleus of spin I, it divides each electronic transition into $2I+1$ equally spaced transitions, all of equal intensity, since all orientations of the nucleus are equally probable. When there is anisotropy, the spacing K in frequency varies according to the formula

$$K^2 g^2 = A^2 g_\parallel^2 \cos^2\theta + B^2 g_\perp^2 \sin^2\theta. \qquad \ldots\ldots(7.17)$$

The spacing ΔH in field varies as K/g, and experimental verification of formulae (7.16) and (7.17) for cobalt is shown by Pryce (1949), to whom these formulae are due.

In general the values of the parameters $g_\parallel$, $g_\perp$, D, A, B and P in the Hamiltonian (7.15) can be obtained immediately from the spectra once the principal axes of

anisotropy have been located. Only relative signs (g is of course always positive as the Hamiltonian has been written) of the parameters can be obtained, and these only from second order effects. Where a fine structure is present, the hyperfine structure varies slightly for the different electronic transitions, and the signs of A and B relative to D can be found (see Bleaney and Ingram 1951 a). For $S=\frac{1}{2}$, if there is an appreciable quadrupole splitting term P, the relative signs of A, B and P can be found using second order shifts combined with those due to the direct effect of the external magnetic field on the nuclear magnetic moment.

The absolute sign of D can be found at low temperatures by two methods which depend on the differing Boltzmann populations of the various states. The electronic splitting produces a small anisotropy in the susceptibility, corresponding to a term in $1/T^2$, whose sign depends on direction relative to the anisotropy axes (see Bleaney and Ingram 1951 a, Griffiths and Owen 1952 a). The second method utilizes the temperature variation of the relative intensities of the different lines of the fine structure (Bleaney, Elliott, Scovil and Trenam 1951).

§ 8. THE IRON GROUP

8.1. *Experimental Results*

The iron group comprises ions with outer electrons in the $3d^n$ configuration, where n runs from 1 to 9. These ions are formed from the atoms in the following table, where the mass and natural abundance of each stable isotope which should give rise to a hyperfine structure (i.e. an isotope for which $I \neq 0$) are also given.

Table 2. The Iron Group

Atomic number	Symbol	Isotopes with $I \neq 0$	
22	Ti	47 (7·5%);	49 (5·5%)
23	V	50 (0·24%);	51 (99·8%)
24	Cr	53 (9·4%);	
25	Mn	55 (100%);	
26	Fe	57 (2·2%);	
27	Co	59 (100%);	
28	Ni	61 (1%);	
29	Cu	63 (69%);	65 (31%)

The significant datum for the paramagnetism is the number of electrons in the 3d shell, and we shall present the results for each configuration separately. In the heading for each is given the spectroscopic state of the free ion, and the value of the spin–orbit coupling constant for each ion, as calculated by Laporte (1928). There is a certain amount of evidence that the values for the ions when bound in the solid state are rather lower. It is assumed that the ion is surrounded by an octahedron of negatively charged ions, giving rise to a crystalline electric potential of predominantly cubic symmetry and with the same sign for all compounds. This is true for nearly all those which have been investigated, and the few exceptions are mentioned in the text. Where a considerable number of compounds of the same ion, with similar properties, have been investigated, only the general trend is given, with references to typical cases. In general the compounds used have belonged to the following groups: the alums (typical formula $K\,R^{\prime\prime\prime}(SO_4)_2 \cdot 12\,H_2O$) with trigonal symmetry for the trivalent ion $R^{\prime\prime\prime}$; the Tutton salts (typified by $K_2R^{\prime\prime}(SO_4)_2 \cdot 6H_2O$) and the fluosilicates ($R^{\prime\prime}SiF_6 \cdot 6\,H_2O$) with respectively tetragonal (or only rhombic) and trigonal symmetry for the

divalent ions R''. An exception to this general method of treatment in the following pages are the cyanides, which are discussed separately in § 8.2 under the heading 'covalent bonding'.

The general characteristics of the resonance phenomenon in the iron group were the subject of a preliminary survey by Bagguley, Bleaney, Griffiths, Penrose and Plumpton (1948). Subsequent results are in substantial agreement with this survey, the most considerable advances lying in the discovery of hyperfine structure, and the detailed analysis of a number of single crystals. A survey of a number of powders of inorganic and organic salts, including work at wavelengths below 1 cm, is given by Lancaster and Gordy (1951, 1952), but the experimental technique has been criticized by Weidner and Whitmer (1952 a).

$3d^1$, $^2D_{3/2}$, Ti^{3+}, $\lambda = 154\ cm^{-1}$. The five-fold orbital level is split by a cubic crystalline field into a doublet and a triplet, the latter being lower. Each orbital level has a twofold spin degeneracy. The salt most investigated is the alum $CsTi(SO_4)_2.12H_2O$, where the crystalline field has also a trigonal component. Under the action of this and the spin–orbit coupling, the sextet (= orbital triplet with twofold spin degeneracy) level is split into three Kramers' doublets, but the separation is not very large (of the order of a few hundred cm^{-1}). At low temperatures only one doublet is populated, with effective $S = \frac{1}{2}$. Owing to the small splitting from the next states, the g-value departs markedly from 2, and is anisotropic, and the spin–lattice relaxation time is abnormally short. Resonance was first observed below 8°K by Bijl (1950 a), using a powder, but the lines were still broadened by the relaxation process. Bogle and Cooke (unpublished), using a single crystal below 4°K, found $g_\parallel = 1{\cdot}25$, and $g_\perp = 1{\cdot}14$. These values are not compatible with the approximate theory of Bleaney (1950 a), but can be fitted when admixtures from the levels at $2\Delta/5$ are included. A larger cubic field splitting of the orbital levels is required ($>50\,000\ cm^{-1}$) than seems likely but this is also true for the chrome alums (see below).

Although titanium has two odd isotopes, of low abundance, no observation of hyperfine structure in the paramagnetic resonance spectrum has yet been reported.

$3d^2$, $^3F_2$, V^{3+}, $\lambda = 104\ cm^{-1}$ (*from spectroscopy*). The sevenfold orbital level is split by a cubic crystalline field as in fig. 6, but with the singlet at the top and the triplets lower. No paramagnetic resonance work has been reported, but Abragam and Pryce (1951 a) give a number of theoretical formulae. The susceptibility of one compound, the ammonium alum, has been investigated by van den Handel and Siegert (1937). Their results indicate that the ground state is a triplet ($S = 1$), with $D = 5{\cdot}0\ cm^{-1}$, $g_\parallel = 1{\cdot}98$, $g_\perp = 1{\cdot}82$. Weak resonance transitions corresponding to transitions between the $M = \pm 1$ levels, with hyperfine structure, should be observable in suitably oriented fields, probably only at low temperatures.

$3d^3$, $^4F_{3/2}$, Cr^{3+}, $\lambda = 87\ cm^{-1}$; V^{2+}, $\lambda = 55\ cm^{-1}$. The splitting of the orbital levels is as in fig. 6, with an orbital singlet lowest, some $10^4\ cm^{-1}$ below the first triplet level. The spin quadruplet remains completely degenerate in a cubic field, but splits into two Kramers' doublets in any field of lower symmetry, the splitting being represented by the terms in D in the spin-Hamiltonian. The effective g-value is close to the free spin value, being given by the formula

$$g = 2 - 72\lambda/5\Delta. \qquad \ldots\ldots(8.1)$$

The chromium salts most investigated are the alums, which were used in much of the early work at room temperature on paramagnetic resonance (Bagguley

and Griffiths 1947, 1950 b, Weiss, Whitmer, Torrey and Hsiang 1947, Whitmer, Weidner, Hsiang and Weiss 1948 a, b, Halliday and Wheatley 1948). This work confirmed the supposition (Van Vleck 1939 a), based on crystallographic evidence, that the four ions in unit cell of the alums are each subject to a trigonal crystalline field whose axis is one of the body diagonals of the cubic unit cell. The separation between the two spin doublets in zero magnetic field varies between 0·12 and 0·18 cm^{-1}, according to the monovalent ion (K, NH_4, Rb, Cs, NH_3CH_3). At lower temperatures Bleaney (1950 b) finds that the three latter alums behave similarly, with slight changes in the values of the splitting, but considerable alterations occur in the first two. The ammonium alum (Bleaney and Penrose 1948) behaves normally, but with a rapidly decreasing splitting, down to 80°K, where the crystals 'shatter', and two splittings appear. The potassium alum shows a similar behaviour down to 160°K, but at lower temperatures the spectrum becomes very complex and the behaviour is not understood. These changes may be connected with the anomalous dielectric behaviour of the alums (Griffiths and Powell 1952).

A diluted chrome alum has been used by Bleaney and Bowers (1951) to establish the nuclear spin of ^{53}Cr as 3/2, with an estimated magnetic moment (see also Bowers 1952) of $(-)0{\cdot}5 \pm 0{\cdot}1$ nuclear magnetons. The hyperfine structure is isotropic ($A=B$), arising practically entirely from a mixture of the $3s3p^63d^34s$ state by configurational interaction. The value of g for the diluted potassium chromium selenate alum used is $1{\cdot}976 \pm 0{\cdot}002$; using $\lambda = 87\ cm^{-1}$, this gives an overall splitting of the orbital levels of $\Delta = 50\,000\ cm^{-1}$, which is almost twice as large as that estimated by Van Vleck (1939 a).

Few salts of divalent vanadium have been investigated. Bleaney, Ingram and Scovil (1951) find that the ammonium Tutton salt, diluted with zinc, has an electronic splitting of about 0·3 cm^{-1} which must be ascribed to a crystalline electric field of rhombic symmetry. An isotropic hyperfine structure for the predominant isotope ^{51}V is observed, with $A = 0{\cdot}0088\ cm^{-1}$.

The nuclear spin of the rare isotope ^{50}V has been shown to be 6 by the method of paramagnetic resonance, independently by Kikuchi, Sirvetz and Cohen (1952), using diluted vanadium ammonium sulphate, and by Baker and Bleaney (1952) using the complex cyanide $K_4V(CN)_6 . 3H_2O$ diluted with the isomorphous diamagnetic ferrous salt. Other results on this and the chromicyanide are given in §8.2.

$3d^1$, $^2D_0$, Cr^{2+}, $\lambda = 57\ cm^{-1}$; Mn^{3+}, $\lambda = 85\ cm^{-1}$. The fivefold orbital degeneracy of these ions is split by a cubic field into a lower doublet and an upper triplet. A tetragonal field splits the doublet, and the ground state has an effective spin $S=2$ equal to the actual spin. Bagguley, Bleaney, Griffiths, Penrose and Plumpton (1948) report resonance at room temperature in chromous oxalate and carbonate. In a diluted Tutton salt $Cr(NH_4)_2(SO_4)_2 . 6H_2O$, Bleaney and Ingram (unpublished) find no resonance, indicating that the spin levels are split, probably by more than 1 cm^{-1} (see Abragam and Pryce 1951 a). No results have been reported for Mn^{3+}, which would be expected to show a hyperfine structure. Measurements on compounds (either of Cr^{2+} or Mn^{3+}) in which the paramagnetic ion is located in a trigonal crystalline field would be interesting, since the lower orbital doublet is not split by such a field. The Jahn–Teller mechanism should then play a marked role, as in $CuSiF_6 . 6H_2O$ (see below).

$3d^5$, $^6S_{5/2}$, Mn^{2+}, Fe^{3+}. As indicated in the theoretical discussion (§7.3) the

case of a half-filled shell (that is, 5 electrons in a d-shell) is a rather special one. With Russell–Saunders coupling an S state results, with no resultant orbital angular momentum. Interaction with the crystalline electric field is then abnormally small, and ions in S states show a susceptibility exceedingly close to the free spin value, following Curie's law down to about 1°K or lower. The important properties from the resonance point of view are thus : g very close to 2·0023; an initial splitting of the levels of the order of a few tenths of a wavenumber; a long spin–lattice relaxation time. Thus conditions are favourable for investigation by resonance methods, and observations at room temperature for some manganese salts (powders) were reported by Cummerow and Halliday as early as 1946.

Detailed measurements on two diluted manganese salts $MnSiF_6 . 6H_2O$ and $Mn(NH_4)_2(SO_4)_2 . 6H_2O$ were made by Bleaney and Ingram (1951 a). These showed that g was very slightly less ($g = 2{\cdot}000 \pm 0{\cdot}001$ for the former salt) than the free-spin value and that the fine structure corresponded to a second degree term $S . D . S$, with a comparatively small cubic term. The second degree term arises mainly from magnetic dipole–dipole interaction when the electron cloud is distorted from spherical symmetry (Pryce 1950 a). A hyperfine structure is also observed, corresponding to the nuclear spin of 5/2. The hyperfine structure is nearly isotropic and nearly the same in the two salts. A similar behaviour, with almost identical hyperfine structure, is found in $Mn SO_4 . 7H_2O$ and manganese acetate (Kumagai, Ono, Hayashi and Kambe 1952).

Of ferric compounds the alums have been most investigated, being easy to grow and commonly used in adiabatic demagnetization work. In the undiluted alums the line width due to magnetic dipole interaction is large, and the spectrum is not well resolved. Nevertheless, the splitting has been shown not to correspond with the cubic field splitting previously assumed in work below 1°K (Bagguley, Bleaney, Griffiths, Penrose and Plumpton 1948, Bijl 1950 b, Weidner *et al.* 1949). Work on diluted iron ammonium alum (Ubbink, Poulis and Gorter 1951, Whitmer and Weidner 1951) indicates that side lines in the fine structure are broadened by inhomogeneity in the crystal field. Meyer (1951) has explained the splitting with considerable success using a combined trigonal and cubic field. These appear to be considerably different in magnitude from values which would fit observations on the undiluted alum. Bleaney and Trenam (1952) find that narrow lines can be obtained with diluted ferric rubidium alum. It appears that the axes of the cubic field are twisted by about 7½° from those of the cubic crystal cell, in agreement with x-ray crystallographic evidence. The constant a of the cubic field splitting (see §7.3) is found to be $-0{\cdot}0134\ cm^{-1}$, which is close to that obtained by Meyer for the ammonium alum. The cubic field splitting is thus considerably greater than that observed in manganese compounds ($a = +0{\cdot}001\ cm^{-1}$ (Bleaney and Ingram 1951 a)). The g-value appears to be slightly greater (2·004) than the free-spin value. Trenam (unpublished), using iron enhanced in the ^{57}Fe isotope, has not observed any hyperfine structure, suggesting that the overall width of any such structure is small (<6 gauss).

$3d^6$, $^5D_4$ Fe^{2+}, $\lambda = -100\ cm^{-1}$. The orbital levels are split in the same way as those for $3d^1$, a cubic field giving an orbital triplet lowest. This, with its fivefold degeneracy in the spin, will be split by the spin–orbit coupling and crystalline fields of lower symmetry. Bagguley, Bleaney *et al.* (1948) report resonance at 1·3 cm wavelength and 20°K in powders of ferrous ammonium and ferrous potassium sulphate. Recent work of Baker and Bleaney (unpublished) on single

crystals has not greatly advanced the position; the absorption lines are weak and comparable in intensity with those of impurities. In ferrous fluosilicate, where the crystal field should have trigonal symmetry, weak lines are observed when the radio-frequency magnetic field is directed along the trigonal axis. This appears to be due to transitions between levels admixed by some mechanism such as the Jahn–Teller effect. These transitions would otherwise be forbidden.

$3d^7$, $^4F_{9/2}$, Co^{2+} $\lambda = -180\,cm^{-1}$. In a cubic field the sevenfold orbital state is split as in fig. 6 but with a triplet state lowest. This triplet, with its fourfold spin degeneracy, is then split into a number of doublets by the combined effect of the spin–orbit coupling and a tetragonal or trigonal field. Resonance has been observed only in the lowest doublet (effective $S = \frac{1}{2}$), and it is found generally that $g_{\parallel}$ lies between 5·5 and 7, and $g_{\perp}$ between 2·5 and 3·5 (Bleaney and Ingram 1949, 1951 b). These values indicate that the spin–orbit coupling is more important than the tetragonal or trigonal field (Pryce 1949). A highly anisotropic hyperfine structure due to ^{59}Co, $I = 7/2$, is observed, with $A \sim 0{\cdot}02 - 0{\cdot}03\,cm^{-1}$, and B varying from $0{\cdot}007\,cm^{-1}$ down almost to zero. The theory is complicated, but the work of Abragam and Pryce (1951 c) is a good agreement with the experimental results. Kambe, Koide and Usui (1952) find also that a reasonable fit can be obtained with measurements of susceptibility, using a less precise theory.

An interesting exception to the above is the blue salt, Cs_3CoCl_5, where the Co^{2+} ion lies in a tetrahedron instead of an octahedron. This gives a cubic field of opposite sign, so that the magnetic behaviour should resemble that of chromium salts, rather than that of other cobalt salts, which are mostly pink. Bagguley, Bleaney *et al.* (1948) found that the powder showed a resonance spectrum not greatly dissimilar from that of the pink salts, but unpublished work of Owen and Stevens suggests that the orbital singlet level is indeed lowest. Within the spin quadruplet resonance is observed only for the $(\frac{1}{2}, -\frac{1}{2})$ transition, and it is probable that the separation between the $\pm 3/2$ level and the $\pm \frac{1}{2}$ level is of the order of $10\,cm^{-1}$.

$3d^8$, $^3F_4$, Ni^{2+}, $\lambda = -335\,cm^{-1}$. The orbital levels in a cubic field split like those of trivalent chromium, leaving an orbital singlet lowest. The effective spin then equals the actual spin $S = 1$, and the g-value is almost isotropic, being given by eqn. (8.1). In a trigonal or tetragonal field the spin triplet splits into a singlet and doublet, but in a rhombic field into three singlets.

The trigonal field case is exemplified by nickel fluosilicate. At room temperature Holden, Kittel and Yager (1949) find a splitting of the spin triplet of $0{\cdot}5_6\,cm^{-1}$, but this falls as the temperature is reduced (Penrose and Stevens 1949) to $0{\cdot}12\,cm^{-1}$ at hydrogen temperatures. g is isotropic at about 2·29. A second compound with a trigonal field is $Ni(BrO_3)_2 . 6H_2O$, which probably has the alum structure. For this Owen (unpublished) finds $g = 2{\cdot}29$, and a splitting of $1{\cdot}93\,cm^{-1}$.

Griffiths and Owen (1952 a) have investigated a number of nickel Tutton salts, mostly at room temperature. They find $g = 2{\cdot}25$ for them all, and that the spin triplet is split into three singlets with an overall separation ranging from $2{\cdot}5\,cm^{-1}$ to $4\,cm^{-1}$. By combining their results with anisotropy measurements they conclude that the smaller splitting in the triplet is usually lower, and that the spin–orbit coupling constant λ is probable nearer $-250\,cm^{-1}$ in the solid. This gives $14\,000\,cm^{-1}$ for the overall cubic field splitting, which is supported by measurements of the optical absorption spectrum.

A certain amount of theoretical work has been carried out on the role played by exchange forces in nickel salts (see § 10).

$3d^9$, $^2D_{5/2}$, Cu^{2+}, $\lambda = -852\,cm^{-1}$ (*Laporte*); the measured value for the free ion is $-828\,cm^{-1}$ (Shenstone, quoted by Abragam and Pryce 1951 c). In a cubic field the orbital levels split into an upper triplet and a lower doublet. In a tetragonal or rhombic field this doublet is further split into two singlets. Resonance is observed for the lower of these, which is doubly degenerate in the spin $S = \frac{1}{2}$. In the theory matrix elements connect this state only with those of the cubic field triplet state, giving approximate g-values $g_\parallel = 2-(4\pm4)\lambda/\Delta$, $g_\perp = 2-(4\mp2)\lambda/\Delta$. Use of the plus or minus signs depends on which orbital singlet is the lower.

The Tutton salts have been much investigated. The concentrated salts (Bleaney, Penrose and Plumpton 1949) all have approximately $g_\parallel = 2{\cdot}4$, $g_\perp = 2{\cdot}1$, but with a rhombic component. The line width shows considerable anisotropy, due to magnetic dipole interaction with two close neighbours, with a secondary contribution from the unresolved hyperfine structure discovered by Penrose (1949). This hyperfine structure is anisotropic, and the theory was not reconciled with the experimental values until the introduction of the promoted s-electron hypothesis (Abragam 1951). Ingram (1949) found that the spectrum showed evidence for a considerable electric quadrupole interaction, and Bleaney, Bowers and Ingram (1951) have analysed this in more detail. For diluted copper potassium sulphate they obtain $A = -0{\cdot}0097\ cm^{-1}$, $B = +0{\cdot}0034\ cm^{-1}$, and $P = +0{\cdot}0010\ cm^{-1}$ for the more abundant isotope ^{63}Cu; for the rarer isotope ^{65}Cu, A and B are 7% larger, and P 8% smaller. The spectrum is shown in fig. 8. The signs were obtained through the combined effect of the quadrupole splitting and the direct interaction of the external field with the nuclear magnetic moment, and are only relative. Later unpublished work shows, however, that there are considerable rhombic components in the g-values and the nuclear interactions.

Copper sulphate and copper acetate also give g-values similar to those of the Tutton salts. These salts show interesting exchange effects, and will be considered later (§10). Preliminary work on three chlorides of copper is reported by Itoh, Fujimoto and Ibamoto (1951). These also show exchange effects, and $CuCl_2.2H_2O$ has been found to become antiferromagnetic below 4·5° K, where its spectrum has been investigated by Ubbink *et al.* (1952).

Bagguley and Griffiths (1952) have found that in some of the copper Tutton salts the g-values and the positions of the axes of anisotropy vary considerably between 90° K and room temperature, $g_\parallel$ becoming smaller than $g_\perp$. This is interpreted as due to a change in the crystal field which brings a different orbital level to the bottom. At the higher temperatures there is considerable broadening by spin–lattice relaxation, and rather surprisingly the characteristic time τ is roughly proportional to $1/T$ above 150° K.

A case of considerable interest is the cupric ion in a trigonal field (as in $CuSiF_6.6H_2O$), since this field does not split the orbital doublet state. Allowing for the spin, the ground state is fourfold degenerate, but this degeneracy is lifted by the Jahn–Teller effect, which distorts the copper–water-molecule complex. There are a continuous set of distortions of the water molecules all giving the same energy (Van Vleck 1939 b), and Abragam and Pryce (1950) have shown that this leads to an isotropic g-value as observed by Bleaney and Ingram (1950). Bleaney and Bowers (1952 a) have since found that the spectrum changes gradually as the temperature is lowered, corresponding ultimately to that of three ions, each in a field of tetragonal symmetry whose

axis is one of the edges of a cube whose trigonal axis is the crystal axis. The effect is thus more complicated than was originally thought, and the theoretical problem has not been fully solved.

8.2. *Covalent Bonding*

The magnetic properties of certain substances which are usually regarded as covalently bonded have been investigated from time to time. As typical of these we shall discuss the cyanides of the iron group elements. The most striking result is that the susceptibilities often differ considerably from those found for the ionically bound ions; for example, the ionic ferric ion shows an effective spin of 5/2 whereas the ferricyanides show spin $\frac{1}{2}$. An explanation of this has been given by Van Vleck (1935) and a detailed treatment has been given by Howard (1935). The basis of the theory amounts to assuming that each electron moves in the smoothed out field of all the other charges, much as is done in the self-consistent field treatment of atoms, except that the smoothed out field is not spherically symmetrical, but approximates to cubic symmetry. This field is sometimes regarded as an electrostatic field which, in a first approximation, replaces the interactions between the electrons. As the charges are moving it is not really an electrostatic field and is usually called the 'intramolecular field'.

If a d-electron is considered as moving in such a field, having cubic symmetry, the fivefold orbital degeneracy is lifted (see fig. 6) and the triplet is lowest. The electrons in the complex go into these levels independently, subject to the exclusion principle. If a modified form of Hund's rule is now used, namely that the electrons are arranged to give the maximum total spin, the ground states and thus the magnetic properties can be determined. This procedure has been followed by Kotani (1949) to discuss the susceptibilities of a number of cyanides. The general agreement is satisfactory. Howard had already concluded that fields of rhombic symmetry exist in the ferricyanides and recent resonance work has confirmed this and shown that the other cyanides also show rhombic symmetry. This means that the theory given by Kotani does not go far enough. The extension has not yet been made, nor is it clear that this is all that is required. Before giving the resonance results it may be pointed out that the case of d^3 is the only one in which the ionic and covalent complexes seem to be alike. This follows directly from the theory, as also does the fact that d^6 should be diamagnetic.

Experimental Results.

The complex cyanides of the iron group which have been investigated belong to the two series $K_3R'''(CN)_6$ and $K_4R''(CN)_6.3H_2O$. In each case the six cyanides form an octahedron (from chemical evidence believed to be nearly regular) round the paramagnetic ion. We shall give the results for each type of ion separately, beginning with the $(3d)^3$ ions (no work is reported on $(3d)^1$ or $(3d)^2$).

$(3d)^3$ ions, Cr^{3+}, V^{2+}. The diluted cyanides of both ions give a g-value of 1·992, which is much closer to the free spin value than is found for the hydrated sulphates of these ions. This suggests that the cubic field is much stronger in the cyanides. An electronic splitting is found in each case; for $K_3Cr(CN)_6$ diluted with cobalticyanide Bowers (1952) finds $D = +0{\cdot}083\,cm^{-1}$, $E = 0{\cdot}011\,cm^{-1}$ and for $K_4V(CN)_6.3H_2O$ Baker and Bleaney (1952) give $D = -0{\cdot}0274\,cm^{-1}$,

$E = 0{\cdot}0072\,\text{cm}^{-1}$. The hyperfine structures due to ^{53}Cr and ^{51}V are about 30% smaller than in the hydrated sulphates.

$(3d)^4$, Mn^{3+}. Although $K_3Mn(CN)_6$ is known to be paramagnetic (though measurements of its susceptibility vary considerably) no resonance was found by Bowers (1952) and no appreciable broadening of resonance lines of Cr^{3+} present as an impurity was observed.

$(3d)^5$, Fe^{3+}, Mn^{2+}. $K_3Fe(CN)_6$ has been investigated by Bleaney and Ingram (1952) who find a single resonance line corresponding to $S = \frac{1}{2}$. The g-values have rhombic symmetry, with extreme values 2·30, 2·18, 0·94. The line widths, which are five to ten times greater than would be expected from pure dipole–dipole interaction, may be due to anisotropic exchange interaction. The line width decreases when the salt is magnetically diluted, showing (together with the lack of temperature dependence) that it is not due to spin–lattice relaxation. The susceptibility departs strongly from Curie's law and can be accounted for if there are excited states a few hundred cm^{-1} higher. Baker (unpublished) finds that $K_4Mn(CN)_6 . 3H_2O$ shows a similar behaviour, with slightly greater anisotropy in its g-values. The hyperfine structure is also anisotropic, as would be expected.

§ 9. THE RARE EARTH GROUP

9.1. *Theory*

In reviewing the theory of the resonance phenomena in rare earth ions the first limitation is that at present the only salts which have been systematically studied are the ethyl sulphates. The theory of these is in a reasonably satisfactory form and we therefore propose to give a short account of its principal features, paying most attention to those aspects which will probably hold for other salts.

The main differences from the iron group ions arise from the following facts:

(1) The magnetic carriers are 4f, rather than 3d electrons.

(2) The symmetry of the crystal field is entirely different and can no longer be regarded as nearly cubic. This is certainly true for the ethyl sulphates and it is known (Penney and Kynch 1939) that the assumption of cubic symmetry will not explain the susceptibilities of the sulphates. Rare earth ions seem, in general, to have nine nearest neighbours which are probably arranged to have a vertical threefold axis of symmetry, with possibly a horizontal plane of reflection. There is a need for further determinations of crystal structures.

(3) The splittings produced by the crystal fields are much smaller than those in the iron group ($\sim 100\,\text{cm}^{-1}$) and are less than those produced by the spin–orbit interactions. This does not imply that the fields are weaker in the crystals but arises because the 4f electrons are well inside the ions and therefore do not feel the full effect of the neighbouring charges. This is partly because the values of $\overline{r^n}$ (see eqn. (6.15)) are smaller than in the iron group and may also be due to screening by the outer electrons.

In developing the theory the same interactions as already listed must be considered, but the order of their treatment is slightly altered, in that the effect of the crystal field must be considered after account has been taken of the spin–orbit interactions. The field is thus taken as a perturbation on degenerate states which are eigenstates of **J**, the total angular momentum. In the ethyl sulphates the crystal field has the form (Elliott and Stevens 1951 a):

$$V = A_2^0(3z^2 - r^2) + A_4^0(35z^4 - 30r^2z^2 + 3r^4) + A_6^0(231z^6 - 315r^2z^4 + 105r^4z^2 - 5r^6) + A_6^6(x^6 - 15x^4y^2 + 15x^2y^4 - y^6) \quad \ldots\ldots(9.1)$$

where all terms with n odd, and greater than 6 have been dropped. The reason for this follows from arguments similar to those already given for iron group ions, and n now ranges up to six because the electrons are f-electrons. The determination of the crystal field matrix elements is in principle no more difficult and the same methods as those described earlier are used (Stevens 1952a). One difference in the results is that with the field (9.1) the electronic levels are always doublets when the number of electrons is odd and singlets or doublets when the number is even. In the latter case a Jahn–Teller distortion is expected and does in fact seem to occur, although the splitting which it produces is very small ($\sim 0{\cdot}05\ \mathrm{cm}^{-1}$). The theory of the nuclear hyperfine structure proceeds in much the same way as in the iron group and when resonance is found the results can be expressed in terms of the usual spin-Hamiltonian (7.15) with $S = \frac{1}{2}$. (The only exceptions are the gadolinium ion, which is in an S-state, and the ions which show the Jahn–Teller effect.) The main complication comes from the fact that with more parameters to describe the fields it is correspondingly more difficult to elucidate their values by working back from the resonance results.

A striking difference between the rare earth g-values and those for the iron group is that whereas in the iron group the majority of g's are near 2, in the rare earths there is a wide variation, and usually considerable anisotropy. This can be seen along the following lines. Suppose we neglect the term in A_6^6 in (9.1) and consider the perturbation on a **J** level. The eigenstates are eigenstates of J_z, by the considerations of §6.3, and $\pm J_z$ are degenerate. In the absence of the crystal field the Landé g-factor (Λ) describes how the levels behave in a field H, and with the field along $0z$ the level which has the component of angular momentum J_z has energy $\beta H \Lambda J_z$. On adding the crystal field the positions of the levels in zero magnetic field are altered, but not the way they diverge when a magnetic field is applied along $0z$. Thus the separation between the levels consisting of the states $\pm J_z$ is $2\beta H \Lambda J_z$, or its g-value is $2\Lambda J_z$. When the field is applied perpendicularly to $0z$, whereas in the free ion the magnetic moment could rotate, it is now tied to the crystal field. In fact the perturbation has no matrix elements between $\pm J_z$ unless $J_z = \pm \frac{1}{2}$. (This is the basis for the selection rules for magnetic dipole transitions quoted earlier, and follows from the theorems of §6.3.) Thus for $J_z \neq \pm \frac{1}{2}$, $g_{\parallel} = 2\Lambda J_z$ and $g_{\perp} = 0$. For $J_z = \pm \frac{1}{2}$ there will also be considerable anisotropy, with $g_{\perp}$ greater than $g_{\parallel}$. If the term in A_6^6 is now included the eigenstates are no longer eigenstates of J_z but consist of admixtures of states which differ by ± 6 in J_z. The g-values are altered and in particular the admixtures between $J_z = \pm 7/2$ and $J_z = \mp 5/2$ now have non-zero $g_{\perp}$. This is important because whenever $g_{\perp}$ is zero no resonances can be observed because of zero transition probabilities (cf. eqn. (5.2)).

Another interesting result is that if the crystal field does not introduce admixtures of states in different J levels it can be proved that $Bg_{\parallel} = Ag_{\perp}$ (Elliott and Stevens, to be published). In erbium ethyl sulphate this relation is closely obeyed and for the other ions it provides a good indication of the approximations which may be made. Experimentally, if this equality holds, the hyperfine structure is isotropic when displayed as a function of field in an experiment carried out at constant frequency. Finally it may be remarked that the nuclear spins of ^{143}Nd, ^{145}Nd, ^{147}Sm, ^{149}Sm and ^{167}Er were first obtained from the paramagnetic resonance spectra, and the theory has been used to estimate the nuclear moments (§9.2).

9.2. *Experimental Results*

In this section we shall give a short survey of the resonance results obtained with ethyl sulphates. In general the relaxation times are much shorter than those found in iron group compounds, so that in several cases it has only been possible to observe resonance by working at the temperatures of liquid helium. This is in accordance with the theory of Van Vleck (1940) since the spins and orbits are closely coupled and the nearest excited levels are usually within 100 cm^{-1}. The published work to which we refer is usually in the form of letters. Full accounts of both the experimental work (H. E. D. Scovil) and the theoretical work (R. J. Elliott and K. W. H. Stevens) are being prepared for publication and will appear in due course. Unless otherwise stated all the results have been obtained on crystals of lanthanum ethyl sulphate with the magnetic ions as impurities, and refer to the ground state.

Cerium Ethyl Sulphate Ce^{3+} $4f^1 {}^2F_{5/2}$. (Experiments: Bogle, Cooke and Whitley 1951. Theory: Elliott and Stevens 1951 b, 1952 b.)

$$g_{\parallel}(5/2) = 3{\cdot}72 \pm 0{\cdot}01 \qquad g_{\parallel}(\tfrac{1}{2}) = 0{\cdot}955 \pm 0{\cdot}005$$

$$g_{\perp}(5/2) = 0{\cdot}2 \pm 0{\cdot}05 \qquad g_{\perp}(\tfrac{1}{2}) = 2{\cdot}185 \pm 0{\cdot}010$$

(These values are more accurate than the published values: private communication from Dr. A. H. Cooke.)

The $\pm\frac{1}{2}$ level is lower by 3 ± 1 cm^{-1}. In the concentrated salt the order of the two levels is reversed. No hyperfine structures have been observed: cerium has no stable odd isotopes.

Neodymium Ethyl Sulphate Nd^{3+} $4f^3 {}^4I_{9/2}$. (Experiments: Bleaney and Scovil 1950, Bleaney, Scovil and Trenam 1952. Theory: Elliott and Stevens 1951 a.)

The level is an admixture of $J_z = \pm 5/2$ with $J_z = \mp 7/2$.

$$g_{\parallel} = 3{\cdot}535 \pm 0{\cdot}001, \quad g_{\perp} = 2{\cdot}073 \pm 0{\cdot}001$$

^{143}Nd $|A| = 380{\cdot}3 \pm 0{\cdot}1$, $|B| = 198{\cdot}9 \pm 0{\cdot}5$, $|P| < 1$

^{145}Nd $|A| = 236{\cdot}4 \pm 0{\cdot}1$, $|B| = 123{\cdot}7 \pm 0{\cdot}5$, $|P| < 1$. Units of 10^{-4} cm^{-1}.

Both isotopes have $I = 7/2$, and $^{143}\mu_N = 1{\cdot}0 \pm 0{\cdot}25$ n.m. and $^{145}\mu_N = 0{\cdot}62 \pm 0{\cdot}15$ n.m. (The values of the nuclear moments are obtained from the more refined theory and are in good agreement with the values given by Murakawa and Ross (1951).)

Samarium Ethyl Sulphate Sm^{3+} $4f^5 {}^6H_{5/2}$. (Experiments: Bogle and Scovil 1952. Theory: Elliott and Stevens 1952 a.)

The level has $J_z = \pm\frac{1}{2}$ with second order admixtures from $J = 7/2$.

$$g_{\parallel} = 0{\cdot}596 \pm 0{\cdot}002, \; g_{\perp} = 0{\cdot}604 \pm 0{\cdot}002$$

^{147}Sm $|A| = 60 \pm 1$, $|B| = 251 \pm 1$, $|P| < 4$

^{149}Sm $|A| = 49 \pm 1$, $|B| = 205 \pm 1$, $|P| < 4$. Units of 10^{-4} cm^{-1}.

Both isotopes have spin 7/2, $^{147}\mu_N = 0{\cdot}83 \pm 0{\cdot}15$ n.m. and $^{149}\mu_N = 0{\cdot}68 \pm 0{\cdot}1$ n.m. (An arithmetical error in the previous values of the nuclear moments has been corrected.)

Gadolinium Ethyl Sulphate Gd^{3+} $4f^7$ ^8S. (Experiments: Bleaney, Elliott, Scovil and Trenam 1951.) The ion is in an S-state and the spin-Hamiltonian is:

$$g\beta H.S + A_2^0\{3S_z^2 - S(S+1)\}$$
$$+ A_4^0\{35S_z^4 - 30S(S+1)S_z^2 + 25S_z^2 - 6S(S+1) + 3S^2(S+1)^2\}$$
$$+ A_6^0\{231S_z^6 - 315S(S+1)S_z^4 + 735S_z^4 + 105S^2(S+1)^2S_z^2$$
$$- 525S(S+1)S_z^2 + 294S_z^2 - 5S^3(S+1)^3 + 40S^2(S+1)^2 - 60S(S+1)\}$$
$$+ \tfrac{1}{2}A_6^6\{(S_+ + iS_y)^6 + (S_z - iS_y)^6$$

where $S = 7/2$, and $g = 1{\cdot}992 \pm 0{\cdot}002$.

3 $A_2^0 = 200{\cdot}3$, 60 $A_4^0 = -3{\cdot}93$, 1260 $A_6^0 = 0{\cdot}52$, 1260 $A_6^6 = -4{\cdot}0$. Units of 10^{-4} cm^{-1}. No hyperfine structure is observed, not because there are no nuclear spins, but, presumably, because of the absence of unpaired s-electrons. Measurements in zero magnetic field show some discrepancies with the above spin-Hamiltonian, though it is not clear why.

Erbium Ethyl Sulphate Er^{3+} $4f^{11}$ $^4I_{15/2}$. (Experiments: Bogle, Duffus and Scovil 1952, Bleaney and Scovil 1951. Theory: Elliott and Stevens 1951 a.)

^{167}Er $g_\parallel = 1{\cdot}47 \pm 0{\cdot}03$, $g_\perp = 8{\cdot}85 \pm 0{\cdot}2$,

$|A| = 52 \pm 1$, $|B| = 314 \pm 1$, $|P| = 30$. Units of 10^{-4} cm^{-1}.

$I = 7/2$, $\mu_N = 0{\cdot}6 \pm 0{\cdot}15$ n.m., $Q \sim 10$ in the units used by Mack (1950). Ground level is probably an admixture of $J_z = \pm 7/2$ with $\mp 5/2$. The resonance g-values are at variance with the optical rotation measurements of Becquerel, de Haas and van den Handel 1937, which suggest a level with $g_\parallel \sim 12$, as also do the susceptibility measurements of van den Handel (1940).

Other Ions. No resonances have been found in dysprosium ethyl sulphate (Dy^{3+} $4f^9$ $^6H_{15/2}$) and in ytterbium ethyl sulphate (Yb^{3+} $4f^{13}$ $^2F_{7/2}$). Theory suggests that this is because the ground states are $J_z = \pm 3/2$ with admixtures which differ by six in J_z. These all have zero $g_\perp$. Resonance has been found in praseodymium (Pr^{3+} $4f^2$ $^3H_4$) by Bleaney and Scovil (1952). This ion has an even number of electrons and the field (9.1) gives a doublet lowest. There is a Jahn–Teller distortion which alters the field but produces only a slight splitting, so that the levels can be regarded as a resolved doublet. The following spin-Hamiltonian has been used to describe the results: $g_\parallel \beta H_z S_z + AS_z I_z + \Delta S_x$ where $g_\parallel = 1{\cdot}69 \pm 0{\cdot}01$, $|A| = 0{\cdot}083 \pm 0{\cdot}001$ cm^{-1}, $|\Delta| \sim 0{\cdot}05$ cm^{-1}. A resonance also occurs in terbium (Tb^{3+} $4f^8$ $^7F_6$).

This completes the survey of the ethyl sulphates, but before leaving the rare earths it might be pointed out that isolated resonances have been reported on other salts, e.g. gadolinium bromate, nitrate and sulphate (Bleaney, Elliott, Scovil and Trenam 1951), cerium sulphates and neodymium sulphate (Kurenev and Salikhov 1951), praseodymium chloride (Davis, Kip and Malvano 1951), cerium oxalate, gadolinium chloride (Lancaster and Gordy 1951). The evidence that the ions which actually give the resonance are those stated is not always convincing, partly because it is very difficult to get rare earth compounds which are free from impurities of other rare earths, and partly because experience with the ethyl sulphates suggests that low temperatures are essential for measurements on ions other than gadolinium.

§ 10. EXCHANGE INTERACTIONS

10.1 *Isotropic and Anisotropic Exchange*

In this section we shall give an account of the effects of the exchange interactions between paramagnetic ions. The usual treatment of the exchange follows that given by Dirac (1947) and consists of showing that the coulomb interactions between electrons can, when all spin–orbit couplings are neglected, be replaced by an interaction between their spins of the form $\frac{1}{4}J(1+4\mathbf{S}_1 . \mathbf{S}_2)$. The constant term is usually neglected as of no interest and the value of J depends on the overlap of the orbital wave functions. In considering the exchange between paramagnetic ions in a crystal this theory can be taken over unchanged when the spin–orbit forces are neglected. Although in many aspects of paramagnetism it is not permissible to neglect the spin–orbit interactions no detailed treatment of the effect on the exchange of its inclusion seems to have been published. The custom is to assume that the interaction is $J\mathbf{S}_1 . \mathbf{S}_2$ where $\mathbf{S}_1$ and $\mathbf{S}_2$ refer to the actual spins of the ions, which are not necessarily $\frac{1}{2}$. This interaction is said to be an 'isotropic exchange interaction' because it depends only on the relative orientation of the spins. It may, however, be noted that the value of J will probably depend not only on the separation but also on the positions of the ions, because, in the crystal field, the charge densities are not spherically symmetrical. We now come to the so-called 'anisotropic exchange'. The isotropic exchange results from the exclusion principle when certain simplifying assumptions are made. If these assumptions are not made there will be some other form of interaction and the term 'anisotropic exchange' is used to account for those parts of this interaction which have a form different from $J\mathbf{S}_1 . \mathbf{S}_2$. If the line joining two ions is an axis of symmetry of the ions and their surroundings the anisotropic exchange can be written in a form similar to that of a dipole–dipole interaction (Ollom and Van Vleck 1951). However for the most part in paramagnetic resonance the exchange has usually been assumed to be purely isotropic.

10.2 *Line Widths*

We now return to the theory of line widths and consider the dipole–dipole and isotropic exchange interactions. The theory of Van Vleck (1948) assumes that the ions behave as free spins, and the method consists of a rigorous calculation of the area and the second and fourth moments of the absorption line. These are then used to discuss the line shapes in a qualitative way. The method of Pryce and Stevens (1950) makes no assumptions about the states of the ions and gives the area and second moment only. The method could be used to obtain the fourth moment but the result would be so complex that it would be of very limited use—the simpler case treated by Van Vleck is already sufficiently complicated. One interesting result from Van Vleck's theory is that there is no contribution to the second moment from isotropic exchange between similar ions, but there is a contribution to the fourth moment. To see the effect of this we may consider the following example. Suppose that in the absence of the exchange the line shape is gaussian. When the exchange is introduced the shape will alter so that its area and second moment are unchanged while its fourth moment is increased. This suggests that the centre part of the line will be narrowed and the excess area distributed in the wings. That is, the line will be peaked. This is the phenomenon known as exchange narrowing. There are, however, other circumstances under which isotropic exchange does not cause narrowing.

Van Vleck's treatment shows that the fourth moment contains the exchange and it seems certain that if the method of Pryce and Stevens were extended to the fourth moment, the exchange would again make a contribution. Van Vleck's second moment, which is obtained on a somewhat unrealistic model of an electronic paramagnetic (it is a much better model for nuclear resonance, for which it was originally deduced) does not contain the exchange between similar ions, whereas when calculations are made using the theory of Pryce and Stevens it usually happens that exchange does make a contribution. It is therefore necessary to decide under what conditions it will broaden, or narrow, a resonance line. We shall do this by deciding the conditions under which it contributes to the mean square width.

The formulae for the second moments are fairly complex expressions. They can however, be given a simple interpretation because it is found that they reduce to summations of the form $\Sigma(W_{ij})^2$, where W_{ij} denotes an interaction between ions i and j. This means that the problem can be broken down, and all that is needed is to take a definite ion and consider the contribution which it gives, taken as interacting with every other ion in turn. That is, the problem reduces to a sum of the results of two-ion problems. Suppose then that a given pair of ions are similar and have effective spin $\frac{1}{2}$. If they precess about parallel axes exchange forces between them do not contribute to the second moment. If however they have different axes of precession the exchange contribution to the second moment depends on $J^2 \sin^4 \frac{1}{2}\theta$ when they have the same precessional frequencies and on $J^2 \cos^2 \theta$ when they have different frequencies, where θ is the angle between the axes. Finally suppose the spins are greater than $\frac{1}{2}$. This usually implies that there are splittings in zero field, and the concept of axes of precession is not easy to formulate. Exchange again gives a contribution. The conclusion is that in nearly all cases exchange does contribute to the second moment—the exception being Van Vleck's case which arises when the ions can be regarded as similar and precessing about parallel axes. Thus in estimating whether exchange will broaden or narrow a resonance line, account must be taken of the relative contributions, or absence of contributions, from the various possible pairs.

Several methods have been used to estimate the magnitude of the exchange interactions. We consider now the direct way of measuring the mean square line widths and subtracting the contribution from dipole–dipole interactions. This method was used by Ishiguro, Kambe and Usui (1951) in the case of nickel fluosilicate. The spin triplet is resolved into a singlet and a doublet in zero field, all the ions are identical and the exchange gives a contribution to the second moment. They assumed that the exchange interactions with other than nearest magnetic neighbours were negligible and found $|J| \sim 0{\cdot}04\ \text{cm}^{-1}$. A similar procedure has been used for nickel ammonium selenate by Stevens (1952b). In this case the spin triplet is completely resolved and it is possible to determine the sign of the exchange. The result is that the sign is that usually associated with ferromagnetism with magnitude of the order of $0{\cdot}025\ \text{cm}^{-1}$. These are the only cases where this method has been used.

10.3. *Copper Sulphate and Acetate*

An interesting phenomenon has been reported by Bagguley and Griffiths (1950a) in copper sulphate. Copper sulphate has two inequivalent magnetic ions in the unit cell (spin $\frac{1}{2}$) and it might be expected that the resonance spectrum would consist of two distinct spectra, one from each type of ion. In fact at 3·04 cm

and 1·23 cm wavelength only one resonance line is observed, and the width of the line varies considerably as the direction of the external field is altered. The explanation of this phenomenon was given by Bagguley, Griffiths and Pryce (1948) and more recently by Yosida (1950) in terms of exchange interactions between dissimilar ions. In the line width discussion it has been pointed out that exchange between similar ions narrows the lines, whereas exchange between dissimilar ions broadens the lines. In this case we have an extreme case of the broadening, the two lines from the different types having coalesced. When this happens the usual treatment of the line widths breaks down and the method used by Pryce was to assume that the exchange energy between dissimilar ions was larger than the differences in their energies in the measuring field. It is then more appropriate to assume that the ions are similar, and to take account of the exchange interactions before introducing the difference energies. The result is that a single resonance line occurs at the position midway between the points at which the separate resonances would occur, and an understanding of the variations of line widths can be obtained from the difference energies. From these considerations it is then clear that if a sufficiently large magnetic field is applied it should be possible to make the difference energies larger than the exchange energies. This experiment was tried by Bagguley and Griffiths by working at 0·85 cm wavelength when, for certain directions of the external field, two lines were obtained at the expected positions. This confirmed the theory and provided an estimate of the magnitude of the exchange between dissimilar ions, giving $|J| \sim 0{\cdot}15\ \text{cm}^{-1}$.

Another interesting effect of exchange interactions has been found by Bleaney and Bowers (1952b) in copper acetate (see also Kumagai, Abe and Shimada (1952)). The crystal structure of this substance is not known (though it is being actively investigated) so that there was no initial knowledge of the number of inequivalent ions in the unit cell. Susceptibility measurements of Guha (1951) had shown an unexpected behaviour and it was of interest to investigate its resonance properties. We shall not give a direct account of these but will instead give an outline of their theoretical interpretation. The conclusion is that the crystal contains pairs of copper ions which are relatively close, and interacting rather strongly together so that they can be considered as more like diatomic 'molecules' than individual ions. Each ion has spin $\frac{1}{2}$, but in the 'molecule' these are coupled together by exchange to give either a triplet spin state or a singlet. The singlet lies lowest and the triplet is some $300\,\text{cm}^{-1}$ higher. Further, anisotropic exchange interactions have split the triplet with splittings of order $0{\cdot}3\ \text{cm}^{-1}$, so that this group behaves rather like the nickel ion in a nickel Tutton salt. The singlet on the other hand is non-magnetic. The result is that at high temperatures all the levels are populated, the resonance and susceptibility results are similar to those for nickel salts; at low temperatures the 'molecules' fall into the singlet state, the susceptibility tends to a constant value and no resonances are found. The exchange between the ions in the pair is antiferromagnetic in sign and of magnitude $300\,\text{cm}^{-1}$. That this is the correct explanation is further confirmed by a study of the hyperfine structure which conforms with that expected for the 'molecules'.

These results, which are rather different from the usual resonance experiments, in which efforts are made to eliminate the interactions, have thus provided direct information about exchange interactions. This is of particular interest in view of recent theories of exchange (Zener 1951).

§ 11. RESONANCE IN GASES AND OTHER SPECIAL COMPOUNDS

The bulk of paramagnetic resonance experiments have been concerned with salts of the iron and rare earth groups, but there are a number of other substances which have been investigated. Chief amongst these are the paramagnetic gases, and organic free radicals. We shall treat these below, mentioning briefly also phosphors and colour centres.

11.1. *Paramagnetic Gases*

Few gases are paramagnetic, since most common molecules have closed electron shells. The best known exceptions to this rule are oxygen and nitric oxide, though there are also a small number of polyatomic gases such as NO_2, ClO_2 and vapours such as $FeCl_3$. So far as we are aware, all resonance work on gases has been carried out by Beringer and Castle, of Yale University.

Before summarizing their results we shall give a brief outline of the notation used to describe diatomic molecules. We begin by neglecting the nuclear spins and assume that there are no forces coupling the rotational motion of the molecular axis with the electronic motions. This enables us to consider the motion when the axis is fixed in space. If we further neglect forces coupling spins with orbits, the total spin **S** is a good quantum number. The total angular momentum **L** is not a good quantum number because the system does not have a centre of symmetry, and we use instead the component of **L** parallel to the axis $0z$, which is a constant. The standard notation is to write $|L_z| = \Lambda$. Then apart from Λ-type doubling the two states L_z and $-L_z$ of **L** have the same energy. States with $\Lambda = 0$, 1, etc. are denoted by Σ, Π, etc., and the spin multiplicity by a superscript, e.g. $^3\Sigma$, $^2\Pi$. Suppose we now include an interaction between spins and orbits. This ties the spin to the molecular axis, $L_z + S_z$ is a good quantum number, and is given as a subscript. Each level will again have twofold degeneracy corresponding to right-handed or left-handed rotation about the axis and typical states might be $^2\Pi_{1/2}$, $^2\Pi_{3/2}$, etc. Including now a coupling between the motion of the axis and the motion relative to the axis it must be remembered that as there are no preferred directions the total angular momentum **J** is conserved and the corresponding level has $(2J+1)$-fold degeneracy. Thus on the twofold degeneracy of $|L_z + S_z|$ is superimposed the degeneracy associated with the rotation of the molecular axis and this is split by the interaction into levels characterized by **J**.

K is used to denote the total orbital angular momentum (molecular plus electronic), so that $\mathbf{K} + \mathbf{S} = \mathbf{J}$.

Oxygen O_2. The oxygen molecule is in a $^3\Sigma$ state, having two unpaired spins and zero orbital momentum about the axis. The ^{16}O nucleus has $I = 0$, so no nuclear effects need be considered. The coupling of the orbital motion with the spin and rotation is small, since $\Lambda = 0$ and the main coupling is between spin and rotation. An orbital moment **K** combines with the spin moment **S**($= 1$) to give levels having $J = K$, $K \pm 1$. The separations of these levels are such that $J = K+1$ and $K-1$ are nearly degenerate and separated by about $2\,cm^{-1}$ from $J = K$. These splittings are not greatly dependent on the value of K (which takes only odd integral values, by symmetry requirements). The physical reason for the splittings is that magnetic dipole interactions between the two electron spins are greater when they point along the axis than when they are perpendicular to it. This separates $J = K$ from $J = K \pm 1$ and the degeneracy of the $J = K \pm 1$

levels is then lifted by the smaller magnetic interaction between the electron spin moment and the field generated by the rotation of the molecule as a whole.

An oscillating magnetic field can induce transitions for which $\Delta J = \pm 1$ and $\Delta K = 0$ (Van Vleck 1947) and, apart from the $J = 0 \leftrightarrow 1$ line these fall near 2 cm^{-1}. They have been observed by the methods of gaseous microwave spectroscopy. If a steady magnetic field is applied there is a Zeeman splitting of the J levels and magnetic transitions can be induced between the components. These have been observed by Beringer and Castle (1951). In a sufficiently small field the Zeeman splitting is linear, and the highest effective g-value is 1, this being for the $K = 1$, $J = 1$ and 2 levels. The magnetic moment is due only to the electron spin, which is fixed, so that the g-values of other levels are lower owing to their greater total angular momenta. This simple model breaks down at the fields used by Beringer and Castle (up to 9 kG and with frequency 0·31 cm^{-1}) since the field begins to break down the internal coupling. As a result a considerable number of the many rotational levels populated at room temperature give resonance lines in the region investigated, some forty lines being resolved. The exact theory is complicated, but Henry (1950) has calculated numerically the positions of a number of transitions and these agree closely with those measured experimentally. The identity of the lines was checked by observing the changes in relative intensities as the temperature was reduced from 300° to 90°K. The half width varied from line to line, but an average value for $K = 1$ lines at 300°K corresponds to 0·051 cm^{-1} at atmospheric pressure.

Nitric Oxide NO. The NO molecule has one unpaired electron and one unit of angular momentum about the axis; the ground level is thus $^2\Pi$. This is split by spin–orbit interactions into two doublets, the lower, $^2\Pi_{1/2}$, being diamagnetic because the projections of the orbital and spin magnetic moments on the axis are equal and opposite, and the upper, $^2\Pi_{3/2}$, being paramagnetic. The separation between the two doublets is 120 cm^{-1}. The upper doublet is then combined with the rotational levels to form states of total angular momenta $\mathbf{J}$. It then happens that only the $J = 3/2$ level has a g-value (4/5) which falls in the experimental range. The field used to observe resonance was strong enough to cause some decoupling, giving a triplet fine structure. Each of these lines is then further subdivided into three components of equal intensity, this being a hyperfine structure due to the nitrogen nucleus (^{14}N) of spin $I = 1$. The hyperfine structure arises from a magnetic interaction of the form $A\mathbf{J}.\mathbf{I}$ and a quadrupole interaction which, for $J = 3/2$ and $I = 1$, can be written $\frac{1}{60}eQq[(15/4) - M^2](2 - 3m^2)$ where q is the gradient of the electric field at the nitrogen nucleus. Although in the experiments the external field was strong enough to decouple I and J, and our usual selection rule $\Delta m = 0$ holds for the nuclear magnetic quantum number, the quadrupole interaction energy here depends on the value of M, and the hyperfine structure is therefore not equally spaced. This permits the determination of the size of both the magnetic and the quadrupole nuclear interactions. Beringer and Castle (1950) find $A = 29{\cdot}8 \pm 0{\cdot}3$ Mc/s ($9{\cdot}94 \pm 0{\cdot}1 \times 10^{-4}$ cm^{-1}), and the quadrupole coupling $eQq = -1{\cdot}7 \pm 0{\cdot}5$ Mc/s ($-0{\cdot}57 \pm 0{\cdot}02 \times 10^{-4}$ cm^{-1}), the negative sign being unambiguously determined. The pressure of NO used was about 1 mm Hg, and the half width at half intensity of the lines corresponded to a value of 0·11 cm^{-1} at 76 cm Hg and 25° C. The theory of the Zeeman effect in NO is given by Margenau and Henry (1950).

Nitrogen Peroxide NO_2. Castle and Beringer (1950) have observed a spectrum of three overlapping lines at about 10 mm Hg pressure, which is partially resolved into a large number of lines at about 1 mm Hg. They interpret this as a Paschen–Back effect on the three magnetic moments in the molecule, that of the electron spin $\Sigma = \frac{1}{2}$, that of the ^{14}N nucleus with $I = 1$, and that due to rotation of the molecule. The latter is an asymmetric top and many rotational levels are occupied; the unresolved triplet is due to hyperfine structure in a large number of rotational levels. The electron-spin–nuclear-spin interaction constant A is 132 Mc/s, and that for the interaction of the spin and the molecular magnetic moment is about half as great.

11.2. *Free Radicals, Phosphors, etc.*

A number of organic compounds containing no metallic ions exhibit paramagnetism, owing to the presence of 'free radicals' with unpaired electron spins. Observation of paramagnetic resonance in such compounds was first reported independently by Holden, Kittel, Merritt and Yager (1949, 1950) and Kosirev and Salekov (1947). As would be expected, the g-value of such compounds (Holden, Yager and Merritt 1951, Townes and Turkevitch 1950) is very close to the free spin values, though generally slightly greater. In some cases the lines are much narrower than would be expected from dipole–dipole coupling, indicating that there are considerable exchange effects. Since the magnetic concentration is relatively high, the sharpness of the lines makes the intensity very high. For this reason diphenyl-trinitro phenyl hydrazyl $(C_6H_5)_2\,N{-}NC_6H_2(NO_2)_3$, which has a half-width of only 1·4 gauss even in powdered form, and a g-value of $2{\cdot}0038 \pm 0{\cdot}0002$, is often used as a convenient standard of comparison when measuring other g-values very close to 2. A small speck placed in the microwave cavity next to the salt under investigation gives a sharp, easily recognized line, whose distance from the line whose g-value is required may be determined from the spectrum displayed on an oscilloscope, or recorded. No auxiliary equipment is required, and errors arising from field inhomogeneity (such as when a proton resonance coil is located some distance from the paramagnetic sample) are avoided.

Recent developments in resonance work with free radicals include the discovery of a small anisotropy in the g-values of a single crystal of the hydrazyl radical, and also in picryl-n-amino carbazyl (Cohen, Kikuchi and Turkevitch 1952). Of especial interest is the observation of hyperfine structure, attributed to the nucleus ^{14}N, in the resonance spectrum of the peroxylamine disulphonate ion $(SO_3)_2\,NO^{2-}$ in solution by Pake, Townsend and Weissman (1952). Other free radicals in solution have been investigated by Hutchinson, Pastor and Kowalsky (1952).

The resonance spectrum of Mn^{2+} has been observed in zinc sulphide phosphor by Schneider and England (1951). An isotropic hyperfine structure with $A = 0{\cdot}0063$ cm^{-1} is observed and a small electronic splitting is found with crystals of the hexagonal (wurtzite) structure, which later work has shown to be absent in the cubic (zinc blende) structure. The same workers have observed the Mn^{2+} spectrum in aqueous solutions (see also Tinkham, Weinstein and Kip 1951). Preliminary work on colour-centres is reported by Hutchinson (1949), Schneider and England (1951), Tinkham and Kip (1951) and Schneider, Day and Stein (1951).

§ 12. CONCLUSION

In this report we have discussed mainly those aspects of paramagnetic resonance which are well understood, and coincide with those parts of the subject which have been most investigated. We have not attempted to give details of every salt whose resonance spectrum has been examined; we understand that such information will appear tabulated in a forthcoming volume of Landolt–Börnstein (*Atom- und Molekularphysik*). We shall conclude by mentioning the fields in which the resonance technique has not yet been widely applied.

The paramagnetic properties of the higher transition groups 4d, 5d and 5f–6d are little understood, and only preliminary resonance work has been reported. Griffiths and Owen (1952b) have begun investigations of the 4d and 5d groups, and they have determined the nuclear spins of both ruthenium 99 and 101 as 5/2, using ruthenium hexamine trichloride. The trivalent ruthenium ion has five electrons in the 4d shell, and its magnetic behaviour seems to resemble that of potassium ferricyanide, with covalent bonding. A structure has also been resolved in a diluted salt of iridium ($5d^5$) and resonance is reported in several compounds of molybdenum ($4d^1$ and $4d^3$). Ramaseshan and Suryan (1951) find that a trivalent molybdenum compound, believed to be K_2MoCl_5, gives a g-value of 1·76. Bowers (unpublished) has investigated a powdered compound of divalent silver ($4d^9$) and finds g to be anisotropic and slightly over 2. This behaviour resembles that of Cu^{2+} ($3d^9$). Baker and Bleaney (unpublished) have investigated a number of compounds of uranium down to 20°K without detecting resonance.

The application of paramagnetic resonance to other types of substances, such as phosphors and free radicals, has been mentioned in § 11, and it seems likely that resonance work will play an increasing role in the fields of chemical and solid state investigations. In magneto-chemistry the magnetic properties of metallic ions have long been used as indications of the nature of their chemical bonding, and an understanding of the resonances in covalent complexes is particularly desirable. It requires a more detailed treatment of the bonding mechanism than has hitherto been applied to this problem, and precise data from magnetic resonance work should help in extending the various treatments which have been proposed.

REFERENCES

Abragam, A., 1951, *Physica*, **17**, 209.

Abragam, A., and Pryce, M. H. L., 1950, *Proc. Phys. Soc.* A, **63**, 409; 1951 a, *Proc. Roy. Soc.* A, **205**, 135; 1951 b, *Ibid.* A, **206**, 173; 1951 c, *Ibid.* A, **206**, 164.

Bagguley, D. M. S., Bleaney, B., Griffiths, J. H. E., Penrose, R. P., and Plumpton, B. I., 1948, *Proc. Phys. Soc.*, **61**, 542, 551.

Bagguley, D. M. S., and Griffiths, J. H. E., 1947, *Nature, Lond.*, **160**, 532; 1948, *Ibid.*, **162**, 538; 1950 a, *Proc. Roy. Soc.* A, **201**, 366; 1950 b, *Ibid.* A, **204**, 188; 1952, *Proc. Phys. Soc.* A, **65**, 594.

Bagguley, D. M. S., Griffiths, J. H. E., and Pryce, M. H. L., 1948, *Nature, Lond.*, **162**, 538.

Baker, J. M., and Bleaney, B., 1952, *Proc. Phys. Soc.* A, **65**, 952.

Becquerel, J., de Haas, W. J., and van den Handel, J., 1937, *Physica*, **4**, 543.

Becquerel, J., and Opechowski, W., 1939, *Physica*, **6**, 1039.

Beringer, R., and Castle, J. G., 1950, *Phys. Rev.*, **78**, 581; 1951, *Ibid.*, **81**, 82.

Bethe, H. A., 1929, *Ann. Phys., Lpz.*, **3**, 133.

Bijl, D., 1950 a, *Proc. Phys. Soc.* A, **63**, 405; 1950 b, *Thesis*, University of Leiden.

BLEANEY, B., 1946–47, *Rep. Prog. Phys.*, **11**, 178 (London : Physical Society) ; 1950 a, *Proc. Phys. Soc.* A, **63**, 407; 1950 b, *Proc. Roy. Soc.* A, **204**, 203 ; 1951 a, *Physica*, **17**, 175; 1951 b, *Phil. Mag.*, **42**, 441.

BLEANEY, B., and BOWERS, K. D., 1951, *Proc. Phys. Soc.* A, **64**, 1135 ; 1952 a, *Ibid.* A, **65**, 667 ; 1952 b, *Proc. Roy. Soc.* A, **214**, 451.

BLEANEY, B., BOWERS, K. D., and INGRAM, D. J. E., 1951, *Proc. Phys. Soc.* A, **64**, 758.

BLEANEY, B., ELLIOTT, R. J., and SCOVIL, H. E. D., 1951, *Proc. Phys. Soc.* A, **64**, 933.

BLEANEY, B., ELLIOTT, R. J., SCOVIL, H. E. D., and TRENAM, R. S., 1951, *Phil. Mag.*, **42**, 1062.

BLEANEY, B., and INGRAM, D. J. E., 1949, *Nature, Lond.*, **164**, 116 ; 1950, *Proc. Phys. Soc.* A, **63**, 408 ; 1951 a, *Proc. Roy. Soc.* A, **205**, 336; 1951 b, *Ibid.* A, **208**, 143; 1952, *Proc. Phys. Soc.* A, **65**, 953.

BLEANEY, B., INGRAM, D. J. E., and SCOVIL, H. E. D., 1951, *Proc. Phys. Soc.* A, **64**, 601.

BLEANEY, B., and PENROSE, R. P., 1948, *Proc. Phys. Soc.*, **60**, 395.

BLEANEY, B., PENROSE, R. P., and PLUMPTON, B. I., 1949, *Proc. Roy. Soc.* A, **198**, 406.

BLEANEY, B., and SCOVIL, H. E. D., 1950, *Proc. Phys. Soc.* A, **63**, 1369 ; 1951, *Ibid.*, **64**, 204 ; 1952, *Phil. Mag.*, **43**, 999.

BLEANEY, B., SCOVIL, H. E. D., and TRENAM, R. S., 1952, *Phil. Mag.*, **43**, 995.

BLEANEY, B., and TRENAM, R. S., 1952, *Proc. Phys. Soc.* A, **65**, 560.

BLOCH, F., HANSEN, W. W., and PACKARD, M., 1946, *Phys. Rev.*, **69**, 127.

BOGLE, G. S., COOKE, A. H., and WHITLEY, S., 1951, *Proc. Phys. Soc.* A, **64**, 931.

BOGLE, G. S., DUFFUS, H. J., and SCOVIL, H. E. D., 1952, *Proc. Phys. Soc.* A, **65**, 760.

BOGLE, G. S., and SCOVIL, H. E. D., 1952, *Proc. Phys. Soc.* A, **65**, 368.

BOWERS, K. D., 1952, *Proc. Phys. Soc.* A, **65**, 860.

CASTLE, J. G., and BERINGER, R., 1950, *Phys. Rev.*, **80**, 114.

COHEN, V. W., KIKUCHI, C., and TURKEVITCH, J., 1952, *Phys. Rev.*, **85**, 379.

CONDON, E. U , and SHORTLEY, G. H., 1951, *The Theory of Atomic Spectra* (Cambridge : University Press).

COOKE, A. H., 1950, *Rep. Prog. Phys.*, **13**, 276 (London : Physical Society).

CUMMEROW, R. L., and HALLIDAY, D., 1946, *Phys. Rev.*, **70**, 433.

DAVIS, C. F., KIP, A. F., and MALVANO, R., 1951, *Atti. Accad. Naz. Lincei*, **11**, 77.

DIRAC, P. A. M., 1947, *The Principles of Quantum Mechanics* (Oxford : University Press).

ELLIOTT, R. J., and STEVENS, K. W. H., 1951 a, *Proc. Phys. Soc.* A, **64**, 205; 1951 b, *Ibid.* A, **64**, 932 ; 1952 a, *Ibid.* A, **65**, 370 ; 1952 b, *Proc. Roy. Soc.* A, **215**, 437.

FERMI, E., 1930, *Z. Phys.*, **60**, 320.

GORDY, W., 1948, *Rev. Mod. Phys.*, **20**, 668.

GORTER, C. J., 1947, *Paramagnetic Relaxation* (Amsterdam : Elsevier) ; 1948, *Physica*, **14**, 504.

GRIFFITHS, J. H. E., and OWEN, J., 1952 a, *Proc. Roy. Soc.* A, **213**, 459; 1952 b, *Proc. Phys. Soc.* A, **65**, 951.

GRIFFITHS, J. H. E., and POWELL, J., 1952, *Proc. Phys. Soc.* A, **65**, 289.

GUHA, B. C., 1951, *Proc. Roy. Soc.* A, **206**, 353.

HALLIDAY, D., and WHEATLEY, J., 1948, *Phys. Rev.*, **74**, 1712.

VAN DEN HANDEL, J., 1940, *Thesis*, University of Leiden.

VAN DEN HANDEL, J., and SIEGERT, A., 1937, *Physica*, **4**, 871.

HENRY, A. F., 1950, *Phys. Rev.*, **80**, 396.

HOLDEN, A. N., KITTEL, C., MERRITT, F. R., and YAGER, W. A., 1949, *Phys, Rev.*, **75**, 1614; 1950, *Ibid.*, **77**, 147.

HOLDEN, A. N., KITTEL, C., and YAGER, W. A., 1949, *Phys, Rev.*, **75**, 1443.

HOLDEN, A. N., YAGER, W. A., and MERRITT, F. R., 1951, *J. Chem. Phys.*, **19**, 1319.

HOWARD, J. B., 1935, *J. Chem. Phys.*, **3**, 813.

HUTCHINSON, C. A., 1949, *Phys, Rev.*, **75**, 1769.

HUTCHINSON, C. A., PASTOR, R. C., and KOWALSKY, A. G., 1952, *J. Chem. Phys.*, **20**, 534.

INGRAM, D. J. E., 1949, *Proc. Phys. Soc.* A, **62**, 664.

ISHIGURO, E., KAMBE, K., and USUI, T., 1951, *Physica*, **17**, 310.

ITOH, J., FUJIMOTO, M., and IBAMOTO, H., 1951, *Phys. Rev.*, **83**, 852.

JAHN, H. A., and TELLER, E., 1937, *Proc. Roy. Soc.* A, **161**, 220.

KAMBE, K., KOIDE, S., and USUI, T., 1952, *Prog. Theor. Phys., Japan*, **7**, 15.

KIKUCHI, C., SIRVETZ, M. H., and COHEN, V. W., 1952, *Phys. Rev.*, **88**, 142.

KITTEL, C., 1949, *Phys. Rev.*, **76**, 743.
KOSIREV, B. M., and SALEKOV, S. G., 1947, *Akad. Nauk Dokl.*, **58**, 1023.
KOTANI, M., 1949, *J. Phys. Soc. Japan*, **4**, 293.
KRAMERS, H. A., 1930, *Proc. Acad. Sci. Amst.*, **33**, 959.
KUMAGAI, H., ABE, H., and SHIMADA, J., 1952, *Phys. Rev.*, **87**, 385.
KUMAGAI, H., ONO, K., and KAMBE, K., 1952, *Phys. Rev.*, **87**, 374.
KURENEV, V. Y., and SALIKHOV, S. G., 1951, *Zh. Eksper. Teor. Fiz.*, **21**, 864.
LANCASTER, F. W., and GORDY, W., 1951, *J. Chem. Phys.*, **19**, 1181 ; 1952, *Ibid.*, **20**, 740.
LANDOLT–BÖRNSTEIN, 1953 or later, *Atom- und Molekularphysik*.
LAPORTE, O., 1928, *Z. Phys.*, **47**, 761.
LEVINTHAL, E., ROGERS, E. H., and OGG, R. A., 1951, *Phys. Rev.*, **83**, 182.
MACK, J. E., 1950, *Rev. Mod. Phys.*, **22**, 64.
MALVANO, R., and PANETTI, M., 1950, *Nuovo Cim.*, **7**, 28.
MARGENAU, H., and HENRY, A., 1950, *Phys. Rev.*, **78**, 587.
MEYER, P. H. E., 1951, *Physica*, **17**, 899.
MURAKAWA, K., and ROSS, J. S., 1951, *Phys. Rev.*, **82**, 967.
OLLOM, J. F., and VAN VLECK, J. H., 1951, *Physica*, **17**, 205.
PAKE, G. E., TOWNSEND J., and WEISSMAN, S. I., 1952, *Phys. Rev.*, **85**, 682.
PENNEY, W. G., and KYNCH, G. J., 1939, *Proc. Roy. Soc.* A, **170**, 112.
PENROSE, R. P., 1949, *Nature, Lond.*, **163**, 992.
PENROSE, R. P., and STEVENS, K. W. H., 1949, *Proc. Phys. Soc.* A, **63**, 29.
PRYCE, M. H. L., 1949, *Nature, Lond.*, **164**, 116 ; 1950 a, *Phys. Rev.*, **80**, 1107 ; 1950 b, *Proc. Phys. Soc.* A, **63**, 25.
PRYCE, M. H. L., and STEVENS, K. W. H., 1950, *Proc. Phys. Soc.* A, **63**, 36.
PURCELL, E. M., TORREY, H. C., and POUND, R. V., 1946, *Phys. Rev.*, **69**, 37.
RAMASESHAN, S., and SURYAN, G., 1951, *Phys. Rev.*, **84**, 593.
SCHNEIDER, E. E., DAY, R J., and STEIN, G., 1951, *Nature, Lond.*, **168**, 645.
SCHNEIDER, E. E., and ENGLAND, T. S., 1951, *Physica*, **17**, 221.
SCHLAPP, R., and PENNEY, W. G., 1932, *Phys. Rev.*, **42**, 666.
SCHLICHTER, C. P., and PURCELL, E. M., 1949, *Phys. Rev.*, **76**, 466.
STEVENS, K. W. H., 1952 a, *Proc. Phys. Soc.* A, **65**, 209 ; 1952 b, *Proc. Roy. Soc.* A, **214**, 237.
TINKHAM, M., and KIP, A. F., 1951, *Phys. Rev.*, **83**, 657.
TINKHAM, M., WEINSTEIN, R., and KIP, A. F., 1951, *Phys. Rev.*, **84**, 848.
TOWNES, C. H., and TURKEVITCH, J., 1950, *Phys. Rev.*, **77**, 148.
UBBINK, J., POULIS, J. A., and GORTER, C. J., 1951, *Physica*, **17**, 213.
UBBINK, J., POULIS, J. A., GERRITSEN, H. J., and GORTER, C. J., 1952, *Physica*, **18**, 361.
VAN VLECK, J. H., 1932, *Electric and Magnetic Susceptibilities* (Oxford : University Press) ; 1935, *J. Chem. Phys.*, **3**, 807 ; 1939 a, *Ibid.*, **7**, 61 ; 1939 b, *Ibid.*, **7**, 72 ; 1940, *Phys. Rev.*, **57**, 426 ; 1947, *Ibid.*, **71**, 413 ; 1948, *Ibid.*, **74**, 1168.
VAN VLECK, J. H., and PENNEY, W. G., 1934, *Phil. Mag.*, **17**, 961.
VAN VLECK, J. H., and WEISSKOPF, V. F., 1945, *Rev. Mod. Phys.*, **17**, 227.
WEIDNER, R. T., WEISS, P. R., WHITMER, C. A., and BLOSSER, D. R., 1949, *Phys. Rev.*, **76**, 1727.
WEIDNER, R. T., and WHITMER, C. A., 1952 a, *J. Chem. Phys.*, **20**, 749 ; 1952 b, *Rev. Sci. Instrum.*, **23**, 75.
WEISS, P. R., WHITMER, C. A., TORREY, H. C., and HSIANG, J. S., 1947, *Phys. Rev.*, **72**, 975.
WHITMER, C. A., and WEIDNER, R. T., 1951, *Phys. Rev.*, **84**, 159.
WHITMER, C. A., WEIDNER, R. T., HSIANG, J. S., and WEISS, P. R., 1948 a, *Phys. Rev.*, **74**, 1478 ; 1948 b, *Ibid.*, **73**, 1468.
YOSIDA, K., 1950, *Prog. Theor. Phys., Japan*, **5**, 1047.
ZENER, C., 1951, *Phys. Rev.*, **81**, 440.

SEMICONDUCTOR CIRCUIT ELEMENTS

By J. S. BLAKEMORE,* A. E. De BARR† and J. B. GUNN†

* Standard Telecommunication Laboratories Ltd.

† Elliott Brothers (London) Ltd., Research Laboratories

CONTENTS

	PAGE
§ 1. Introduction	161
§ 2. Electrons in solids	161
2.1. Band theory of solids	161
2.2. Impurities in semiconductors	164
2.3. Semiconductor statistics	165
2.4. Surface effects	169
§ 3. Bulk properties of semiconductors	171
3.1. Electrical properties	171
3.1.1. Conductivity	171
3.1.2. Minority carriers	174
3.1.3. Effect of a magnetic field	177
3.1.4. Noise	178
3.1.5. Thermoelectric power	179
3.2. Optical properties	180
3.3. Nucleon-irradiated semiconductors	181
3.4. Preparation of semiconductors	182
§ 4. Metal–semiconductor contacts and p–n junctions	184
4.1. Historical	184
4.2. Metal–semiconductor contacts	184
4.3. p–n junctions	188
4.4. Transistor action	190
§ 5. Experimental results and applications	192
5.1. Properties of point contacts	192
5.1.1. Forward characteristic	192
5.1.2. Inverse characteristic	196
5.1.3. Signal properties	200
5.2. Preparation of p–n junctions in germanium	201
5.3. Properties of p–n junctions	202
5.3.1. Forward characteristic	203
5.3.2. Inverse characteristic	203
5.3.3. Signal properties	205
5.4. Technological applications	205
5.4.1. Diodes	205
5.4.2. Photocells and counters	208
5.4.3. Transistors	208
5.4.4. Other circuit elements	212
References	213

ABSTRACT. Recent advances in the theory of the conduction mechanism in semiconductors and semiconductor–metal systems are reviewed, together with their properties and applications. Since the more important advances relate to systems whose properties depend upon the simultaneous flow of electrons and holes, discussion is limited mainly to those materials, such as silicon and germanium, in which this phenomenon is observed. A brief account is given of the fundamental concepts of semiconductor theory after which such bulk properties of semiconductors as conductivity, behaviour of minority carriers, magnetic effects, noise and optical effects are discussed. Modern theories of rectification and transistor action at metal–semiconductor contacts and p–n junctions are then considered, and descriptions given of the mechanisms and properties of practical circuit elements such as crystal diodes and transistors.

§1. INTRODUCTION

RESEARCH on the properties of semiconducting materials has been greatly stimulated by two important discoveries: (*a*) the observation (Benzer 1944) that point contact rectifiers made from germanium could withstand very high inverse voltages, and (*b*) the discovery (Bardeen and Brattain 1948) of transistor action in germanium. Efforts to understand these phenomena have led to further discoveries of academic and technological importance, and in both fields progress has been very rapid. The aim of this Report is to present an account of the physical phenomena underlying the operation of those semiconductor crystal devices (crystal rectifiers, transistors, etc.) whose properties depend on the presence of charge carriers of both signs. Semiconductor circuit elements such as thermistors, copper oxide rectifiers, selenium rectifiers and photocells and silicon carbide resistors, which do not clearly fall into this category, are not considered especially.

Section 2 is a short account of the theoretical treatment of semiconducting materials and discusses the important concept of the Fermi level. This is followed in §3 by a brief review of the properties of semiconducting materials. In §4 thephenomena at the junction between a metal and a semiconductor or between two semiconductors are considered, and §5 deals with the properties of the various semiconductor crystal devices.

Most of the important results of the kind considered here relate to the elements of Group IV of the periodic classification, and this Report will refer mainly to these elements and, in particular, to germanium and silicon, but the physical explanations given can apply also to other semiconducting materials though the actual details may differ slightly.

§2. ELECTRONS IN SOLIDS

2.1. *Band Theory of Solids*

A satisfactory theory of semiconductors emerges as one consequence of the band theory of solids, which was developed primarily in order to explain metallic behaviour. The various stages in the development of the band theory have recently been discussed in these Reports by Raynor (1952), and will only be briefly reviewed here.

The first step towards the establishment of a theory of metallic conduction was made by Drude (1900), who suggested that the outer electrons of the atoms in a metal might form a 'free electron gas', in thermal equilibrium with the solid. The members of this gas would move freely through the solid, subject to scattering by the atoms of the lattice, and would all participate in conduction. This idea was further explored by Lorentz (1905, 1909) on the assumption that the electrons obeyed Maxwell–Boltzmann statistics, and his results enjoyed a measure of agreement with experiment, notably with regard to the prediction of the Wiedemann–Franz ratio.

A serious objection to the early theory was the prediction of a large electronic specific heat for metals, which is not observed in practice. This difficulty was resolved by the application of quantum mechanics to the problem. Using Fermi–Dirac statistics, based on the Pauli exclusion principle, Sommerfeld (1928) showed that the energies of electrons should be restricted to a number of closely spaced 'levels', each of which can be occupied by only one electron. On Sommerfeld's model, the density of levels at energy ϵ for unit volume of metal, is given by

$$g(\epsilon)\,d\epsilon = 4\pi(2m_n/h^2)^{3/2}(\epsilon-\epsilon_0)^{1/2}\,d\epsilon \qquad \ldots\ldots(2.1)$$

where m_n is a quantity called the 'effective mass' of a free electron—usually of the same order of magnitude as the normal electronic mass—and ϵ_0 is the energy of the lowest permitted level (apart from those of the tightly bound electrons in the inner electronic shells of each atom). The parabolic form of $g(\epsilon)$ as a function of energy is shown in fig. 1.

At the absolute zero of temperature every permissible level is occupied up to an energy ϕ, called the Fermi level, whilst every level at a higher energy is empty. At higher temperatures, some of the electrons in levels slightly below ϕ become thermally excited to higher levels, the degree of population at any energy being given by

$$f(\epsilon)=[1+\exp\{(\epsilon-\phi)/kT\}]^{-1}. \qquad \ldots\ldots(2.2)$$

This is appreciably different from either zero or unity only for energies within a few times kT of the Fermi level. Thus since $\phi-\epsilon_0 \gg kT$ for metals at normal temperatures, only a small fraction of the nominal free electron concentration is free to participate in energy changes and contribute towards electrical conductivity and specific heat. Such an electron gas is said to be degenerate, since the electronic specific heat is degenerated considerably from the classical value of $3k/2$ per electron.

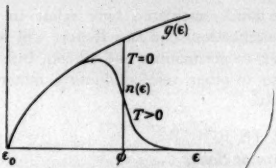


Fig. 1. Distribution of permitted electron levels $g(\epsilon)$ and of electrons $n(\epsilon)$ in these levels as a function of energy, on the Sommerfeld free electron theory.

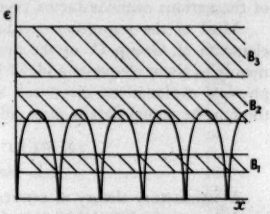


Fig. 2. Permitted energy 'bands' and the periodic potential energy in a crystal lattice.

The Sommerfeld theory gives an adequate explanation of the properties of some simple metals, notably the alkalis, but fails in more complex systems. This is a result of neglecting the internal electric field in a solid, which arises from the periodic array of atomic cores, and from the other electrons in the gas. The way in which this field must be taken into account depends on whether the electrons are to be regarded as fairly tightly bound to the parent atoms, or whether the potential field is weak. The result is similar in both cases, and it is found (e.g. Bloch 1928) that the wave function of an electron in a solid is

$$\Psi=\exp(\pm ikr)\,.\,U_k(r) \qquad \ldots\ldots(2.3)$$

where $U_k(r)$ has the periodicity of the lattice, and k is the wave number, related to momentum.

The substitution of this function into Schrödinger's equation yields real values of k only for certain ranges of electron energy. A plot in three dimensional k-space of the critical values of k at which energy discontinuities occur yields a set of concentric polyhedra, centred at the origin of k-space. The regions between polyhedra, in which energy can vary continuously with wave number, are the Brillouin zones.

If the crystal is made up of M unit cells then, taking account of the two possibilities for the spin, the permitted range of energy in each Brillouin zone comprises $2M$ energy levels and these levels lie in 'bands' separated by forbidden energy gaps in which no electron levels normally exist.

The band system of a solid arises naturally from the discrete energy levels of an isolated atom which become broadened through interaction between the closely spaced atoms of a solid in such a way that the width of the energy band is independent of the number of atoms in the crystal, but depends upon their separation, and on the energy level from which the band arises. If there is more than one atom per unit cell, each energy level may split into several overlapping bands, each requiring a Brillouin zone to represent it. The critical energies at zone boundaries depend on crystallographic orientation, and the highest possible energy for one direction in one zone may exceed the lowest possible energy for a different direction in the next zone. Thus the energy bands B_1, B_2, B_3, etc. shown in fig. 2 may correspond to one or more zones. The complete electronic structure of a solid may bear little resemblance to the original set of levels in the single atom, but the essential feature of permitted and forbidden regions of energy remains.

The chance of an electron near one atom passing to the next depends on the height and width of the interatomic potential barrier relative to the energy of the electron. Thus electrons in the higher (conduction) bands are free to move throughout the lattice, whereas those in the lower (valence) bands are more closely identified with particular atoms.

The distinction between metals and insulators is readily apparent on the band scheme, since the number of electron levels in each band is limited. The available electrons fill the lowest bands completely, and the distribution of electrons in higher bands will be determined by the Fermi–Dirac probability function (2.2). If the number of electrons per atom is such that the highest occupied band is only partly full, then metallic conduction will be found, for electrons with energies near to the Fermi level can be accelerated by an external electric field, thus rising to slightly higher, previously unoccupied, levels in the band. On the other hand, if the number of electrons per atom is just sufficient to fill the highest occupied band, then electrons in this band can only exchange energies with each other. Since all electrons are identical, this does not contribute towards electrical conductivity, and such a material is an insulator. Conduction is possible only if electrons can be excited across the forbidden energy gap into the conduction band. At ordinary temperatures in most semiconducting materials, only about one electron in every 10^{15} is thermally excited in this way. Excitation into the conduction band is also possible if an electron, in passing from one atom to another, can acquire sufficient energy from an applied field to jump the gap (Zener 1934). Since the width of the energy gap is usually of the order of one electron volt, excitation in this way is possible only with fields of about 10^8 v/m.

At high temperatures, however, some electrons are thermally excited across the forbidden energy gap into the empty conduction band above. This is the situation in an intrinsic semiconductor (Wilson 1931 a), and results in an electrical conductivity which increases rapidly as temperature rises. This increase arises not only from the electrons excited into the conduction band, but also from the presence of vacant levels in the valence band, which allow electrons in this band to be accelerated and retarded by electric fields.

The position of a vacant level in the valence band can move about by the breaking of an adjacent valence bond, the ejected electron filling the original vacancy, which has therefore been moved to another site. Under the influence of an electric field the random motion of the vacant bond acquires a systematic component, and it moves as though current were carried by a particle popularly known as a 'hole', with a positive charge numerically equal to that of an electron and an effective mass m_p. The hole concept plays a very important part in semiconductor problems, and Shockley (1951 a) has recently discussed very fully the analogy between the wave mechanical properties of holes and electrons in solids.

The number of semiconductors in which definite evidence of intrinsic conduction has been established is small, since appreciable thermal excitation will only be possible when the intrinsic gap is not unduly large compared with kT. This condition is satisfied at suitable temperatures for germanium (Lark-Horovitz *et al.* 1946), silicon (Pearson and Bardeen 1949), and lead sulphide (Chasmar and Putley 1951) amongst other materials, but most non-metals melt or decompose before the temperature is reached at which intrinsic conduction might be expected to become appreciable.

2.2. *Impurities in Semiconductors*

Most of the known semiconductors derive their finite conductivity from the presence of small quantities of suitable impurities. In these extrinsic conductors, it is believed (Wilson 1931 b) that impurities can give rise to permitted electron levels in the forbidden energy gap. These are localized levels, which do not interact with each other appreciably, since the spacing between impurity centres is much greater than the normal interatomic distance for all but the most impure semiconductors. Thus conduction by passage of electrons from one impurity atom to the next should be negligible, except perhaps at very low temperatures (Hung 1950, Erginsoy 1950 b).

The conductivity of an extrinsic conductor devolves from the thermal excitation of current carriers from the impurity levels to the conduction or valence bands, where they can take part in electrical conduction. If the impurity levels are occupied at zero temperature, and have an energy $\delta \ll \epsilon_i$ lower than the bottom of the conduction band, as in fig. 3, then these electrons can be excited into the empty band at much lower temperatures than those required for appreciable intrinsic excitation. Impurity levels which are normally occupied at low temperatures are known as donors, and the type of conduction produced by their ionization is termed n-type. These materials are identified experimentally by the negative Hall and thermoelectric coefficients, which show that current is carried by electrons.

In p-type semiconductors, on the other hand, 'acceptor' impurity levels, which are empty at absolute zero, are found near to the valence band. At higher temperatures, electrons are thermally excited into these levels, and the resultant holes in the valence band give a mode of conduction characterized by positive Hall and thermoelectric coefficients, indicating that the effective current carriers are positively charged.

Impurity centres can arise from both foreign atoms and lattice defects. Thus phosphorus, with five electrons in its outer shell, acts as a donor impurity in silicon (Pearson and Bardeen 1949), which requires four electrons to complete

the valence bonds of its tetrahedral crystal lattice, whilst boron, with only three outer electrons, is an acceptor impurity in the same semiconductor.

In many compound semiconductors, the lattice defects consequent upon deviations from stoichiometric proportions form impurity centres. Donors arise from an excess of the electro-positive constituent, either in the form of interstitial ions (e.g. ZnO) or vacant electro-negative ion sites (e.g. the alkali halides). In either case, an electron is bound fairly weakly to the singularity, and can be thermally excited to the conduction band.

Vacant electro-positive ion sites form the acceptor centres in a number of p-type compounds (e.g. NiO). In this case, an electron is normally missing from a neighbouring atom, in order to preserve electrical neutrality. This constitutes a 'bound hole', and can be thermally excited to move through the lattice.

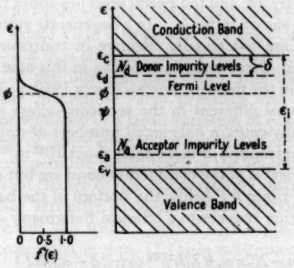


Fig. 3. Band structure of a semiconductor showing possible positions of donor and acceptor levels. The Fermi level ϕ may occur at any point on the energy scale as the criterion of electrical neutrality.

2.3. *Semiconductor Statistics*

The way in which the various electron levels in a semiconductor are filled at a given temperature depends, as for a metal, on the spread of the Fermi–Dirac probability function

$$f(\epsilon) = [1 + \exp\{(\epsilon - \phi)/kT\}]^{-1}. \qquad \ldots\ldots(2.4)$$

The Fermi level ϕ has not the same obvious significance for a semiconductor as for a metal, since in the latter case ϕ is the energy of levels which are just half filled, whilst in a semiconductor ϕ is frequently in a part of the energy scale where no permitted levels exist. Even so, ϕ must act as a normalizing parameter, varying with temperature in such a way that the integral of occupied levels over the whole energy scale is equal to the number of electrons in the crystal. At a given temperature, the number of free carriers for a semiconductor in thermal equilibrium is uniquely defined by ϕ, though the same situation might be reached through a variety of combinations of different impurities.

The concentrations of free electrons in the conduction band, n, and of free holes in the valence band, p, can be expressed in terms of the Fermi energy by making appropriate assumptions about the distribution in energy of levels within these bands. It is usually assumed that the distribution in the conduction band follows the parabolic relationship of the Sommerfeld free electron model:

$$g(\epsilon) = 4\pi(2m_n/h^2)^{3/2}(\epsilon - \epsilon_c)^{1/2}. \qquad \ldots\ldots(2.5)$$

For a real solid, this is not accurately true over a large range of energy, deviations being ascribed to a variation of effective carrier mass with energy. Fortunately, this effect is negligible for semiconductors unless very large free electron concentrations are involved.

The free electron concentration must be the sum of all occupied levels from the bottom of the conduction band upwards. The upper limit of integration is unimportant provided that it is sufficiently high above the Fermi level to ensure that nearly all the free electrons have lower energies. For convenience, infinite energy can be set as the upper limit.

Thus
$$n=\int_{\epsilon_c}^{\infty} g(\epsilon)f(\epsilon)\,d\epsilon. \qquad \ldots\ldots(2.6)$$

When n is very large ($>10^{25}$ m^{-3}) the lower levels in the conduction band are almost completely occupied, and the Fermi level lies above the bottom of the band, as in the metallic case shown in fig. 1. In a degenerate system such as this, all the conduction electrons are not equally effective in contributing towards the conductivity. The Fermi–Dirac integral (2.6) can in this case be represented by the analytical forms used in the theory of metals (Seitz 1940, p. 149).

The situation is very different in the semiconductors used in most practical devices. These usually contain only small numbers of impurities, insufficient to produce carrier concentrations of more than about 10^{24} m^{-3} at room temperature. Thus even the lowest levels of the conduction band are but sparsely filled, and the Fermi level is several times kT below the bottom of the band. In this case, the Fermi–Dirac distribution reduces to a classical Boltzmann distribution, and n has the form
$$n=2(2\pi m_n kT/h^2)^{3/2}\exp\{(\phi-\epsilon_0)/kT\} \qquad \ldots\ldots(2.7)$$
so that the conduction electron concentration in a 'classical' semiconductor would be unchanged if all the distributed levels of the conduction band were replaced by $N_c=2(2\pi m_n kT/h^2)^{3/2}$ levels per cubic metre placed at $\epsilon=\epsilon_c$. This classical limit, whereby all free electrons are equally effective in determining the electrical properties, will be valid for all the semiconductors we shall be concerned with in this Report.

By an argument similar to the above, assuming that the density of levels near the top of the valence band increases parabolically as energy decreases, it can easily be shown that
$$p=N_v\exp\{(\epsilon_v-\phi)/kT\} \qquad \ldots\ldots(2.8)$$
where $N_v=2(2\pi m_p kT/h^2)^{3/2}$ has the same significance for the valence band as N_c has for the conduction band.

For an intrinsic semiconductor, where impurities have negligible effect, n and p must be equal. Equating (2.7) and (2.8) to the intrinsic concentration n_i, we have
$$\begin{aligned} n_i&=(N_cN_v)^{1/2}\exp\{(\epsilon_v-\epsilon_0)/2kT\}\\ &=(N_cN_v)^{1/2}\ \exp(-\epsilon_i/2kT). \end{aligned} \qquad \ldots\ldots(2.9)$$

Since $\epsilon_i \gg kT$, the exponential term dominates the temperature dependence of (2.9); and a semi-logarithmic plot of n_i against $1/T$ is almost linear. A slight curvature results from the temperature dependence of N_c and N_v.

The Fermi level in an intrinsic case is situated at the energy
$$\psi=\tfrac{1}{2}(\epsilon_c+\epsilon_v)-\tfrac{1}{2}kT\ln(N_c/N_v). \qquad \ldots\ldots(2.10)$$

The ratio $N_c/N_v=(m_n/m_p)^{3/2}$ is never far from unity, so that the temperature dependence of ψ is small, and the level is always near to the midpoint in energy between the two bands, being exactly halfway if $m_n=m_p$.

Shockley (1949, 1951 a) has treated $-\psi/e$ as an effective electrostatic potential of the semiconductor. The significance of choosing the energy ψ for this purpose is that when impurities are introduced, changing both n and p, then simple and symmetrical expressions obtain for these quantities in terms of the shift of the Fermi level from ψ. Provided that neither carrier concentration is sufficiently large for degeneracy to be significant, it follows that

$$n=n_1\exp\{(\phi-\psi)/kT\} \qquad \ldots\ldots(2.11)$$

and

$$p=n_1\exp\{(\psi-\phi)/kT\}. \qquad \ldots\ldots(2.12)$$

Thus $np=n_i^2$ for any non-degenerate semiconductor, and

$$np/T^3=4(2\pi k/h^2)^3(m_nm_p)^{3/2}\exp(-\epsilon_i/kT). \qquad \ldots\ldots(2.13)$$

It might therefore be expected that a semi-logarithmic plot of np/T^3 against $1/T$ for a semiconductor would yield a slope and intercept corresponding with eqn. (2.13). The slope $-\epsilon_i/k$ thus obtained usually agrees with the results of independent measurements of the intrinsic gap width, but the intercept may differ considerably from the theoretical value (Pearson and Bardeen 1949). This discrepancy is one consequence of a small variation of ϵ_i with temperature, which results from the change of lattice spacing with temperature (Bardeen and Shockley 1950), and from the temperature dependence of the amplitude of lattice vibrations (Fan 1951). The actual change in ϵ_i is comparatively small (a few parts in 10^4 per degree), and does not seriously invalidate most properties calculated on the assumption of a constant gap width. But if ϵ_i is replaced by $\epsilon_i-\xi T$, then the coefficient of the exponential in eqn. (2.13) is multiplied by a factor $\exp(\xi/k)$. Pearson and Bardeen found experimentally that this multiplying factor was 34 for silicon, which gives a value for ξ of 3×10^{-4} ev/°c, in fairly good agreement with theory.

When appreciable quantities of active impurity are present, one of the carrier concentrations becomes negligibly small compared with the other. Thus in an n-type semiconductor, p is usually many orders of magnitude smaller than n, and only the latter requires consideration for conduction calculations.

The procedure for deriving n and the corresponding position of ϕ when specific amounts of impurity are present was laid down by Wilson (1931 b) for the simple case in which only one type of impurity is present. These procedures have since been discussed for a wider variety of the cases met with in practice (Hutner, Rittner and du Pré 1950, Blakemore 1952).

When a semiconductor is considered in which N_d donor levels per cubic metre are situated at an energy ϵ_d, which is rather lower than the conduction band, as in fig. 3, then these will usually be occupied by

$$n_d=\frac{N_d}{1+\exp\{(\epsilon_d-\phi)/kT\}} \qquad \ldots\ldots(2.14)$$

electrons per cubic metre, whilst the concentration of empty states is N_d-n_d. But in the absence of acceptor impurities, the latter quantity must be the free electron concentration, if the valence band is far below the Fermi level, and is completely full. By equating N_d-n_d to $N_c\exp\{(\phi-\epsilon_c)/kT\}$, which also determines the position of ϕ for a given free electron concentration, independent expressions can be obtained for n and ϕ.

At low temperatures, such that only a small proportion of the donors is ionized (reserve conditions), it is found that

$$n=(N_dN_c)^{1/2}\exp\{(\epsilon_d-\epsilon_c)/2kT\}=(N_dN_c)^{1/2}\exp(-\delta/2kT) \qquad \ldots\ldots(2.15)$$

and that the Fermi level in this case is

$$\phi=\tfrac{1}{2}(\epsilon_d+\epsilon_c)-\tfrac{1}{2}kT\ln(N_c/N_d) \qquad \ldots\ldots(2.16)$$

which lies near to the midpoint of energy between the conduction band and the impurity levels. These equations are very similar in form to eqns. (2.9) and (2.10) for an intrinsic semiconductor : thus it is important to verify by independent methods whether an exponential variation of n with $1/T$ for a semiconductor is the result of intrinsic or extrinsic excitation.

When an appreciable fraction of the donors is ionized, (2.15) and (2.16) have to be replaced by more complicated expressions. The Fermi level is coincident with the donor energy when these are half ionized, and is lower still for more complete ionization.

When ϕ is several times kT below the donor levels these are almost completely ionized, i.e. $n=N_d$. In this exhaustion range, the Fermi level is given by

$$\phi=\epsilon_c-kT\ln(N_c/N_d) \qquad \ldots\ldots(2.17)$$

and falls twice as rapidly with rising temperature as in the reserve region. Exhaustion conditions are found at normal temperatures in such semiconductors as germanium and silicon, for which the activation energy δ of chemical impurities is small. On the other hand, many compound semiconductors such as oxides have impurity activation energies of up to an electron volt, and these maintain reserve conditions up to high temperatures.

The Fermi level in an exhaustion semiconductor cannot fall indefinitely as temperature rises, and eventually a position is reached where appreciable intrinsic excitation occurs. ϕ is then nearing the limiting value ψ for an intrinsic semiconductor, halfway across the intrinsic gap. This situation occurs at about 100°c for germanium containing moderate amounts of impurity, and a radical change occurs in the electrical properties above this temperature. For less pure samples, the influence of impurities persists up to rather higher temperatures.

Arguments exactly similar to those used for n-type semiconductors are applicable in p-type cases, and the resultant expressions for free hole concentration, etc., are identical in form. This results from the symmetrical form of the Fermi–Dirac probability for energies above and below ϕ, and the opposite signs of electron and hole energies. Thus ϕ rises with temperature towards the intrinsic value as acceptors become exhausted in a p-type semiconductor.

In any real solid, it is impossible to avoid the presence of both donors and acceptors (though it is usually ensured that one concentration is as small as possible). If an n-type material contains N_d donors, and a smaller number N_a of acceptors, then each of the latter will trap an electron, leaving only N_d-N_a electrons available for conduction. This can lead to some very interesting effects when N_d and N_a are comparable in magnitude.

All the expressions given above refer to thermal equilibrium conditions in a semiconductor, and the Fermi level has no real significance when the equilibrium distribution is disturbed. A slight disturbance does in fact occur whenever an external electric field is applied, though the effect is very small for

moderate field strengths. However, a more serious disturbance of thermal equilibrium occurs when carriers of anomalous sign are injected (see §3.1.2). For instance, if numbers of holes in excess of the equilibrium value are injected into n-type material, equal numbers of free electrons will arrive within a very short time to restore neutrality, and both n and p will be larger than their usual values. The two concentrations will tend to decay towards the normal values by recombination of holes with electrons, but until this process is complete, the normal Fermi level has no significance.

In order to deal with this situation, Shockley (1949, 1951 a) has proposed the use of non-equilibrium 'quasi-Fermi' levels, ϕ_n and ϕ_p, which govern the concentrations of carriers in the conduction and valence bands respectively by expressions similar to (2.11) and (2.12) for a semiconductor in equilibrium. We define ϕ_n and ϕ_p by

$$n = n_1 \exp\{(\phi_n - \psi)/kT\} \qquad \ldots\ldots(2.18)$$

and

$$p = n_1 \exp\{(\psi - \phi_p)/kT\}. \qquad \ldots\ldots(2.19)$$

As the injected carriers decay, ϕ_n and ϕ_p both approach the thermal equilibrium value of ϕ. It will be seen in §4 of this report that use of the concept of quasi-Fermi levels can assist in the understanding of rectification and transistor action.

The concentrations of carriers in a semiconductor can only be found experimentally by correctly interpreting properties such as electrical conductivity, thermo-power, Hall constant, etc. The problems involved in analysing the bulk behaviour of semiconductors are discussed in §3.

2.4. *Surface Effects*

Many of the practical applications of semiconductors involve the emission of current carriers into free space, or into another solid. Thus it is important to consider the position of the electronic levels of a solid with reference to the energy of an electron outside, and the ways in which the energy differences can be modified.

A convenient energy to regard as zero is that of an electron at rest in free space far from any charged body. Since electrons are not spontaneously emitted in large numbers from solids at normal temperatures, it is obvious that a considerable potential barrier impedes the escape of electrons from solids, and that the conduction band of a semiconductor lies far below the vacuum energy level, as shown in fig. 4(*a*). In diagrams of this kind, energy diagrams, the lower the energy level, the less the potential energy of the electron and therefore the greater the energy required to remove it. The application of an electric field causes the energy levels to become tilted—fig. 4(*b*). Electrons, tending to reduce their potential energy, flow downhill on the energy diagram whilst holes tend to flow uphill.

The shape of the energy levels may also be affected by space charges in the semiconductor. This is a consequence of the condition that, for equilibrium, the Fermi level must be constant throughout any system in electrical contact. Thus, in n-type germanium, in p-type inclusions or in regions in which the density of n-type impurity centres is less than average, the Fermi level (2.17) lies further below the bottom of the conduction band than in the rest of the material. In order that the Fermi level shall be constant, the energy levels must be distorted as shown in fig. 4(*c*). Electrons are drawn in to the abnormal region from the surrounding semiconductor until the space charge distribution due to their (negative) charge

and the positive space charge in the regions from which they have been withdrawn, gives rise to a field which opposes this process and restores equilibrium. A potential barrier is produced which impedes the flow of both holes and electrons.

In fig. 4 (*a*) the electron affinity $\chi = -\epsilon_c$ is the minimum energy required to remove an electron from the bottom of the conduction band, and $\Phi = -\phi$ is the thermionic work function. In a non-degenerate case, Φ is larger than χ by the amount ζ, which may have a value up to $\epsilon_i/2$ for an n-type semiconductor, and is larger than this for a p-type material. Electrons in the conduction band which receive energies less than χ may attempt to leave the semiconductor; but are attracted back in by an 'image force' effect, though this can be partially overcome by application of an accelerating electric field. This reduction of the effective work function by an electric field is the Schottky emission effect.

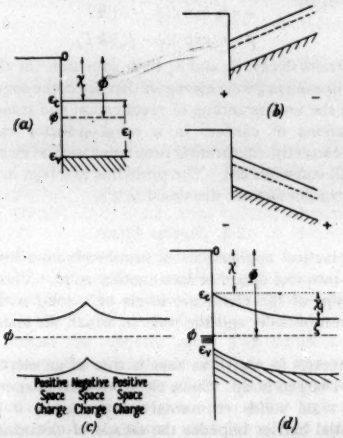


Fig. 4. Energy level diagrams. (*a*) Potential barrier at the surface of a semiconductor. (*b*) Effect of applied electric field. (*c*) Internal potential barrier. (*d*) Surface states.

Electron emission can also be increased considerably by placing another solid in intimate contact with the surface. The barrier layer theories of rectification at a metal–semiconductor contact (Mott 1939, Schottky 1939) show that the potential barrier which electrons have to surmount in order to pass from one solid to the other is approximately the difference of the two thermionic work functions, which can obviously be made large or small by suitable choice of materials.

These theories, and their applicability, are discussed in §3. It is now recognized that under some conditions the barrier height may depart considerably from the difference of work functions (Meyerhof 1947), and that the work function of a free semiconductor surface may be much less dependent on temperature than should be the case if ζ varies with temperature in the usual way whilst χ remains constant (Smith 1949). These discrepancies can be attributed to the presence of additional electronic states at a crystal surface (Bardeen 1947).

Tamm (1932) showed that surface states with energies below the conduction band can arise naturally at crystal surfaces, as a consequence of the abrupt termination of the periodic potential field in the crystal. The density of these states is likely to be increased considerably by the presence of surface imperfections and foreign atoms, notably adsorbed gases such as oxygen. If these levels are initially unoccupied, or only partly full, then they will tend to trap electrons from the bulk of the semiconductor. The resultant positive space charge in the semiconductor depresses the electron levels in the interior until the Fermi level in the interior equals that in the surface levels, as shown in fig. 4 (*d*).

The depleted layer at the surface is very similar to the depletion layer produced at a metal–semiconductor contact on Schottky's theory of rectification. Thus if a metal is brought near to a surface with a large density of surface states, the further adjustment of energies is very slight, and the barrier height, and hence rectification efficiency, is virtually independent of the work function of the metal used. Moreover, if the surface state density is very high, the Fermi level at the surface will be only slightly dependent on the number of electrons in them, and hence on temperature. In consequence, whilst ζ will increase with temperature, χ_1 will vary in sympathy, and the thermionic work function $\Phi = \chi + \chi_1 + \zeta$ will be almost independent of temperature, as is found for some germanium and silicon surfaces (Smith 1949).

It will be noticed that the Fermi level in the surface region is drawn in fig. 4 (*d*) nearer to the valence band than to the conduction band. Thus the free hole concentration is larger than the free electron concentration for the surface region in this case. It is this feature which enables holes to be injected into the n-type body of a semiconductor by application of an electric field from a biased metal contact. More efficient hole injection can, however, be produced with p–n junctions (§§ 4 and 5).

It would be highly satisfying to be in a position to state that the concept of surface states clarified all the interesting properties of semiconductor surfaces. Unfortunately, this is not the case, and considerable doubt surrounds the role of these states for p-type semiconductors. It is not easy to see what form of surface states would have the necessary electro-positive character for these materials, and it is to be hoped that further investigations may shed some light on this subject.

§ 3. BULK PROPERTIES OF SEMICONDUCTORS

The Fermi level in semiconductors is not fixed as in metals but may vary with temperature, impurity concentration etc. As a result of this the behaviour of electrons in semiconductors differs in several ways from their behaviour in metallic conductors, and it is these differences which make possible the many interesting and important applications of semiconductors as circuit elements.

In this section the bulk properties of semiconductors are surveyed briefly in preparation for the more detailed discussions of actual devices in §§ 4 and 5.

3.1. *Electrical Properties*

3.1.1. *Conductivity*

Some of the properties of the elements of Group IV of the Periodic Classification are given in table 1, which shows the gradual transition from an insulator (diamond), through the semiconductors silicon, germanium and gray tin, to metallic lead as the width of the forbidden gap ϵ_1 decreases. Whilst in this respect the distinction between metals and semiconductors is quite clear, in practice, nevertheless, the effects of temperature and impurity concentration result in properties which

differ only by degree. For instance, at very low temperatures a sufficiently pure semiconductor may have so few free carriers as to be an insulator, whilst a very impure semiconductor will show metallic properties.

Table 1

	Carbon (diamond)	Silicon	Germanium	Gray tin	Lead
Atomic number	6	14	32	50	82
Length of edge of unit cell (Å)	3·56	5·42	5·62	6·46	4·92
Density (kg m^{-3})	3510	2400	5360	5750	11340
Number of atoms (10^{28} m^{-3})	17·7	5·0	4·52	2·96	3·35
Dielectric constant	6·2	12	16		
Mobility of electrons (m^2 $volt^{-1}$ sec^{-1})	0·01–0·1	0·12	Hall 0·26 Drift 0·36	~0·3	
Mobility of holes (m^2 $volt^{-1}$ sec^{-1})	0·2	0·025	0·17	~0·3	
Forbidden energy gap (ϵ_i) (ev)	6–7	1·11	0·76	0·1	Over lapping
Energy to ionize donor, δ (ev)	0·35	0·054	$< 0·04$		
Energy to ionize acceptor (ev)	0·35	0·08	$< 0·04$		

A rigorous analysis of electronic conduction in solids involves consideration of the distribution of quantum states in the Brillouin zones and the changes in distribution caused by applied fields. However, it may be shown (Shockley 1951 a) that for many purposes electrons and holes can be treated as classical particles with charge $\mp e$ having definite positions and velocities and effective masses m_n and m_p which depend on the band structure but are generally of the same order of magnitude as the mass of an electron m. Although the properties of positive holes in the valence band can usually be deduced simply from those of electrons in the conduction band, it is not always possible to give a simple physical picture of the effects involved and it must be remembered that the justification for this treatment can only be provided by the rigorous quantum mechanical theory of the effects.

Typical curves showing the electrical conductivity of an impurity semiconductor as a function of temperature are shown in fig. 5. The curves are all of the same shape and, in general terms, may be accounted for as follows. At very low temperatures the conductivity is small and (for an n-type semiconductor) is due almost entirely to electrons thermally excited from the impurity levels into the conduction band (fig. 4). Since the intrinsic gap width ϵ_i is usually much greater than δ, the number of electrons thermally excited from the valence band at these temperatures will be very small. As the temperature is raised the number of electrons able to participate in electrical conduction increases approximately as $\exp(-\delta/2kT)$, eqn. (2.15), and more rapidly than the resistance to their motion increases with temperature in this region. As the temperature is increased still further the conductivity continues to increase until all the available impurity atoms are ionized. At the same time the number of intrinsic charge carriers (holes + electrons) increases. When all the impurity atoms are ionized—exhaustion range—the conductivity increases less rapidly with temperature and may actually decrease owing to increased scattering. At still higher temperatures the number of intrinsic carriers increases rapidly until they carry most of the current. In this intrinsic range, the conductivity increases as $\exp(-\epsilon_i/2kT)$, eqn. (2.9), and is almost independent of the impurity content. These changes of conductivity,

when due allowance is made for the resistance to motion of the electrons, reflect exactly the fall of the Fermi level with temperature (see Putley 1949). In a metal, of course, the position of the Fermi level, and therefore the number of electrons participating in electrical conduction, is almost independent of temperature. At very low temperatures electrons in impurity levels may contribute to the conductivity of semiconductors. An empirical relationship between the density of carriers and the room temperature resistivity of germanium is shown in fig. 6.

With a given density n of carriers—electrons, say—the observed conductivity will be determined by their mean free path. This is dependent on the scattering of the electrons by (*a*) thermal vibrations of atoms on the semiconductor lattice, (*b*) impurity atoms (Erginsoy 1950a), (*c*) ionized impurity centres (Conwell

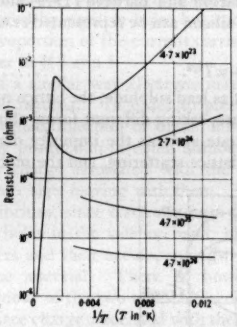


Fig. 5. Resistivity of silicon–phosphorus alloys as a function of temperature (*after* Pearson and Bardeen). The numbers on the curves refer to the number of phosphorus atoms per m³ of melt.

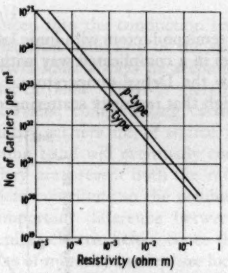


Fig. 6. Empirical relationship between density of carriers and room temperature resistivity of germanium (*after* Taylor).

and Weisskopf 1950). The conduction electrons are in thermal motion and, if the mean free time between collisions is $\bar{\tau}$, the average velocity acquired in an applied field E is given by

$$\bar{v} = \frac{-e\bar{\tau}}{m_n} E \qquad \ldots\ldots(3.1)$$

and the resultant current density is

$$i = -(ne)\bar{v} = \frac{ne^2\bar{\tau}}{m_n} E. \qquad \ldots\ldots(3.2)$$

It is convenient to consider the conductivity in terms of the mobility μ of the carriers, defined as the drift velocity in unit applied field. Thus

$$\mu_n = -e\bar{\tau}/m_n \qquad \ldots\ldots(3.3)$$

and the conductivity σ is $\sigma_n = i/E = ne\mu_n$. $\ldots\ldots(3.4)$

Similarly for holes $\sigma_p = pe\mu_p$ $\ldots\ldots(3.5)$

and in the general case $\sigma = e(n\mu_n + p\mu_p)$. $\ldots\ldots(3.6)$

A complete treatment of conductivity in germanium, including thermal and impurity scattering, has been given by Shockley (1951a).

The general character of the variation of mobility with temperature for an atomic latticed semiconductor is shown in fig. 7. At high temperatures, the mobility is limited only by lattice scattering, and is almost independent of the impurity content. For this process, the mean free path is constant and the corresponding mobility varies as $T^{-3/2}$. The rising trend of mobility as temperature falls is checked by the increasing importance of scattering by impurities, and at low temperatures the limiting process consists of the deflection of the drifting electrons by the electrostatic field of charged impurity atoms (Rutherford scattering). At even lower temperatures, neutral impurity scattering may become important. The mean free time due to ionized impurity scattering varies in a complicated way with temperature, and leads to a mobility which increases approximately as $T^{3/2}$. Pearson and Bardeen (1949) show that mobility data for non-degenerate samples of silicon can be represented reasonably well by an expression of the form

$$1/\mu = uT^{-3/2} + vT^{3/2}. \qquad \ldots\ldots(3.7)$$

For semiconductors with ionic lattices, such as lead sulphide, the lattice mobility varies in a complicated way with temperature, taking different forms above and below the Debye temperature. In degenerate systems the impurity density is so high that impurity scattering dominates lattice scattering, and the mobility is

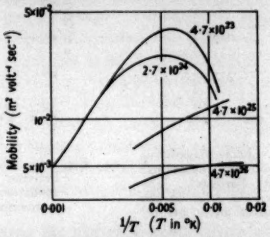


Fig. 7. Mobility of electrons in silicon–phosphorus alloys (*after* Pearson and Bardeen). The numbers on the curves refer to the number of phosphorus atoms per m³ of melt.

constant over a wide temperature range. Theoretical analysis of this case (Johnson and Lark-Horovitz 1947) suggests that the mobility should decrease with increasing impurity content but this is not observed (Pearson and Bardeen 1949).

It has been shown (Ryder and Shockley 1951) that the mobilities of electrons in n-type germanium are independent of field up to fields of about 10^5 v/m but vary inversely as the field thereafter. The effect has been considered by Shockley (1951 b) and the decrease in mobility shown to be due to the influence of scattering by optical modes and to increases of electron temperature. In high fields the drift velocity tends to a constant value, an effect which may be important in interpreting the performance of metal semiconductor contacts (Gunn 1952).

3.1.2. *Minority Carriers*

The normal process of electrical conduction in semiconductors is similar to that in metals. The current carriers are accelerated in an applied field, the effect of their space charge being that any disturbance is propagated with the velocity of propagation of electromagnetic radiation. These so-called normal

carriers are, in n-type semiconductors, electrons moving in the conduction band, and in p-type materials they are positive holes in the valence band. However, Shive (1949a) showed that current could also be carried in semiconductors by carriers of opposite sign to those normally encountered and this important discovery was the starting point for the explanation of the properties of many semiconductor devices. It has been shown that in an n-type semiconductor, electrons may, in certain circumstances, be removed from the valence band, leaving behind positive holes. The process, known as hole injection, can take place at junctions between a metal and a semiconductor or between two semiconductors of opposite conductivity type, and the term 'emitter' contact has been used to describe a contact at which hole injection takes place. At almost all contacts at which current flows into n-type material a fraction of this current is carried by positive holes but special forming treatments are used to increase the proportion of the current carried in this way. The effect will be considered further in §§ 4 and 5.

In a similar way electrons may be introduced into the conduction band in p-type material. In order to preserve the electrical neutrality of the material as a whole, injection of these 'minority' carriers must always be accompanied by the production of equal numbers of normal carriers. There is therefore no space charge associated with the minority carriers, compensating carriers of opposite sign moving with them. These minority carriers are, of course, not in equilibrium, since electrons from the conduction band will eventually combine with holes in the valence band, but whilst they are present both the minority carriers and their accompanying normal carriers contribute to the conductivity of the material. There is, however, an important difference between the behaviour of the two types of carrier in an applied electric field. Since there is no space charge associated with them, the effects of minority carriers are localized and they affect the conductivity of the material only in their immediate vicinity. Under the influence of the applied field they drift with a velocity determined by their mobility until eventually they recombine. In addition to this systematic drift of minority carriers in an electric field, there is always superposed a random spreading due to thermal motion. Even in the absence of an applied field electrons and holes will be in random motion, but normal carriers will, in general, be uniformly distributed so that this random motion produces no external effects. Minority carriers are localized, however, and since each carrier will undergo random diffusion independently of the others, any group of such carriers tends to spread out.

The total minority carrier current arises from superposing the conduction and diffusion current densities, the latter being directly proportional to the gradient of minority carrier density, ∇p or ∇n.

Thus, for example, the total hole current density in n-type material in an applied field is given by

$$i_p = q\mu_p pE - qD_p \nabla p \qquad \ldots\ldots(3.8)$$

where q is the charge on a carrier and D_p is the hole diffusion coefficient, given by the Einstein relation

$$D_p = \boldsymbol{k}T\mu_p/q. \qquad \ldots\ldots(3.9)$$

(3.8) may be expressed in terms of the quasi-Fermi level defined in (2.19) since

$$\nabla p = p\nabla\left(\frac{\psi - \phi_p}{\boldsymbol{k}T}\right). \qquad \ldots\ldots(3.10)$$

Thus

$$i_p = \mu_p p\nabla[\psi - \boldsymbol{k}T \log (p/p_i)] = \mu_p p\nabla\phi_p. \qquad \ldots\ldots(3.11)$$

The flow of electrons and holes in semiconductors has been investigated in detail by Van Roosbroeck (1950) and Prim (1951), who have analysed the flow fields which are set up.

There are two special cases of interest. In the first, the dominant term in (3.11) is the one containing $\nabla\psi$, whilst in the second case, current flow is controlled mainly by the value of $\nabla(\log p)$.

The first case is that of potential flow, where the carriers follow the lines of current flow, and the effects of diffusion can be treated as a perturbation. For example, if $p \ll n_0$ in n-type material, a particularly simple case results, and the flow lines are those of the undisturbed electric field E. A hole signal, such as a δ-function pulse injected at a field source, will split up along the various lines of electric field, and the time t_1 taken to reach a collecting electrode or 'sink' is given by

$$t_1 = \frac{1}{\mu_p}\int \frac{1}{E}\,ds \qquad \ldots\ldots(3.12)$$

where the integration is along a field line. Unless t_1 is independent of the flow line ('isochronous' conditions), the original sharp pulse becomes spread out as a consequence of the range of delay times. It can easily be shown that all fields involving straight lines of flow are isochronous, but further generalization is difficult.

The effect of diffusion is to spread out a sharp pulse into a gaussian shape, of half-amplitude length $3{\cdot}3\,(D_p t)^{1/2}$, where t is the time after injection. Thus a pulse arriving at the 'sink' will have a time length of $3{\cdot}3\,(D_p t_1)^{1/2}/\mu_p E$ seconds, in addition to any spreading resulting from a distribution of delay times.

In diffusion flow, when $\nabla\psi$ is small, the carriers follow the lines of steepest concentration gradient. The flow is formally equivalent to that in problems of isotropic heat conduction. For example, in one-dimensional geometry, a δ-function pulse at the origin will have a distribution of the form $(\pi D_p t)^{-1/2}\exp(-x^2/4D_p t)$ after a time t. Thus a collector at a distance X from an emitter will receive a pulse of carriers at a time $t_2 = X^2/2D_p$ after injection.

When the density of minority carriers is comparable with that of the normal carriers some interesting effects are observed and a discussion of some of these is given by Herring (1949). A small number of holes injected into an n-type semiconductor will set up an electric field which will draw in electrons to neutralize the charge of the holes. However the electron density is so much larger than the hole density that the required field need alter the electron distribution only slightly, and to a first approximation the holes can be considered to move in the same field as would exist if no holes were present. But when the hole and electron densities are comparable, the velocity with which the hole density is propagated in the applied field approaches zero. The increased velocity for low hole densities causes the low density regions to travel faster than those of higher densities. This effect may be observed when pulses of holes are propagated in germanium. The trailing edge of the pulse tends to become very steep—a shock wave phenomenon.

The passage of a group of minority carriers along a semiconductor under the influence of an applied field may be detected by the increased local conductivity of the material as they pass through. In this way the mobility of the injected carriers may be measured and their behaviour studied (Haynes and Shockley 1949). Although the minority carriers eventually disappear by recombination with normal carriers, a hole–electron pair—or 'exciton'—behaves rather like the electron

and proton of an hydrogen atom. In order to recombine it must give up energy and it has been calculated that if this energy is radiated as light quanta the minority carriers should have lifetimes of the order of one second. In practice hole concentrations in n-type germanium are observed to decay exponentially with time constants up to about 200 microseconds. This indicates that recombination accompanied by the emission of light quanta is a relatively unimportant process and that the majority of the holes recombine by exchange of energy with the thermal vibrations of the crystal lattice.

Experiments show that the observed lifetimes increase as the dimensions of the sample are increased and it is concluded that the recombination process in germanium often takes place largely on the surface. Recombination may also occur at grain boundaries and other lattice discontinuities. If the observed lifetime is τ and τ_v is the lifetime due to recombination in the bulk material only then it may be shown that

$$\frac{1}{\tau} = \frac{1}{\tau_v} + \frac{1}{\tau_s} \qquad \ldots\ldots(3.13)$$

where τ_s is the lifetime due to surface recombination. τ_s is conveniently expressed (Shockley 1951 a) in terms of the geometry of the sample and a surface recombination velocity, v_s, which is very dependent upon surface treatment and may vary between $10^{-1} - 10^2$ m/sec for germanium.

3.1.3. *Effect of a Magnetic Field*

When a magnetic field is applied perpendicular to the direction of current flow in a conductor it exerts a force on the electrons, and tends to deflect them sideways, and to set up a space charge along the sides of the conductor. This space charge gives rise to a transverse field of such a magnitude as to counteract the sideways deflection caused by the magnetic field. This is the Hall effect, and the Hall coefficient R is defined as the potential gradient produced by unit applied magnetic field when unit current density flows in the specimen.

For electrons in semiconductors at temperatures so high that thermal scattering is the predominant scattering mechanism

$$R = \frac{3\pi}{8}\frac{1}{ne}\ \mathrm{m^3/coulomb} \qquad \ldots\ldots(3.14)$$

and is positive when the carriers are holes and negative for electrons. When both electrons and holes are present, the Hall constant is a complicated average of the Hall constants and conductivities for each type of carrier. If both thermal scattering and impurity scattering are important, different values for the numerical constant in (3.14) are obtained (Johnson and Lark-Horovitz 1950).

The Hall effect is important because, in semiconductors containing carriers of one sign alone, its sign indicates immediately whether the material is p- or n-type, and also because, from the Hall coefficient and the measured conductivity, it is possible to determine both the concentration and the mobility of the carriers

$$\mu = 8\sigma R/3\pi. \qquad \ldots\ldots(3.15)$$

Mobilities may also be measured by observing the drift velocity, in an applied electric field, of minority carriers injected into the semiconductor. The drift mobility, defined as the ratio of average drift velocity to applied field, is usually not equal to the mobility determined from Hall effect measurements (Hall mobility). If the material contains energy levels in which holes may be trapped so

that they are free to move for only a fraction of the time, then the drift mobility will be only a fraction of the mobility of an untrapped carrier. Such trapping effects will not, however, affect the Hall mobility. In special circumstances (Pearson, Struthers and Theuerer 1949, 1950) mobility may be determined directly from measurements of conductivity. The quantity determined in this way from the conductivity and the total density of excess electrons should be the same as the drift mobility. In n-type germanium the Hall mobility is less than the drift mobility and it has been suggested by Shockley (1951 a) that this may be explained in terms of re-entrant energy surfaces in the Brillouin zone.

Closely related to the Hall effect is the phenomenon of magneto-resistance. The transverse electric field set up in the material prevents the deflection, by the magnetic field, of those electrons having some average velocity. Electrons with velocities greater or less than this average value are deflected by the magnetic field and so suffer increased scattering which is manifested as an increase in resistance (Estermann and Foner 1950). The effect can be calculated on the assumption of spherical energy surfaces in the Brillouin zone, but the observed resistance changes are much greater than can be accounted for in this way. This, together with the fact that the magneto-resistance of single crystals of germanium is strongly anisotropic (Shockley 1950 a), suggests that a satisfactory explanation of this effect also requires a consideration of non-spherical energy surfaces.

Also related to the Hall effect is the Suhl effect (Suhl and Shockley 1949) in which the sideways deflection of holes and electrons moving in a magnetic field may be observed directly. Holes injected into an n-type semiconductor will be deflected towards the same side as the electrons, since the sideways thrust depends upon the product of the velocity and the charge of the particle, both these quantities being reversed for a hole compared with an electron. But the transverse electric field which counteracts the deflection of the electrons increases that of the holes, and their deflection may be observed by measuring the local change of resistance. In large magnetic fields the holes are forced to the surface of the sample, where they recombine more readily than in the interior, so that the decay of hole concentration is greatly increased by the presence of the magnetic field.

3.1.4. *Noise*

When current flows in a circuit containing semiconducting material it is found that, in addition to thermal agitation noise (Johnson noise) which is independent of current, and shot noise which is proportional to current, there is also a contribution to the noise voltage which is proportional to current and which increases rapidly as the frequency is reduced. (The spectral density of both Johnson noise and shot noise is independent of frequency.) Since practical circuits of this kind invariably contain either metal–semiconductor or semiconductor–semiconductor contacts it is difficult to determine whether this flicker noise, as it is called, is due to the properties of the semiconductor itself or to those of the contacts. The same type of noise is also observed in thermionic emission from oxide coated cathodes.

Van der Ziel (1950) has shown that it is impossible for the noise intensity to vary inversely proportionally to the frequency f in the whole interval $0<f<\infty$, but shows that a noise intensity which is inversely proportional to the frequency in a wide range can be obtained if a distribution of correlation times is assumed.

In the case of fluctuating quantities which are caused by a large number of independent events occurring at random, the correlation time is the duration of the event: in the case of fluctuations involving decay problems, the correlation time measures the average decay time. A mechanism which gives rise to a noise spectrum of the type observed has been proposed by Macfarlane (1947). He suggests that flicker noise arises from the diffusion of clusters of impurity centres on to the emitting surface at a relatively small number of points. Emission is assumed to take place only from these points, so that the active area of the surface changes with time. In a later paper (Macfarlane 1950) he suggests that random fluctuations in the concentration of mobile adsorbed ions may modulate the height of the potential barrier at the semiconductor surface. Again emission is assumed to take place only at localized patches on the surface, and the fluctuations in the height of the barrier at these patches give rise to fluctuations in emission.

Gisolf (1949) has investigated the current fluctuations to be expected when an electron is alternately bound and free, and able to contribute to the current only during the time in which it is free—a state of affairs similar to that in a semiconductor. At high frequencies the resulting fluctuations are inversely proportional to frequency and are in reasonable agreement with the observed noise in crystal detectors. A detailed account of noise in crystal rectifiers is given by Torrey and Whitmer (1948).

Montgomery and Shockley (1950) suggest that flicker noise, which is equivalent to a fluctuating resistance, is due to a fluctuating current of minority carriers. They postulate regions in n-type germanium which have a higher hole density than other regions. These may be p-type inclusions or regions which are less strongly n-type than their surroundings. In the presence of an applied field these high densities will be swept away. Holes emitted at the places concerned, in an attempt to re-establish the hole density, will also be swept away, so that the regions behave as sources of holes. The noise produced by a given fluctuating source of holes will be proportional to the lifetime of the holes. If the lifetime is decreased by the application of a magnetic field (Suhl effect) the noise in the semiconductor has been observed to be reduced. Again it is suggested that the noise originates in sources of holes located on the surfaces of the material.

If the hole current were purely random the noise power per cycle would be independent of frequency but it has been suggested (Shockley 1951 a) that certain types of imperfection may give rise to a frequency spectrum of the $1/f$ type.

3.1.5. *Thermoelectric Power*

When a metal is in contact with an n-type semiconductor, equilibrium is reached when the height of the bottom of the conduction band in the interior of the semiconductor above the Fermi level in the metal is approximately

$$\zeta = -kT \log \frac{nh^3}{2(2\pi m_n kT)^{3/2}}. \qquad \ldots\ldots(3.16)$$

The thermoelectric power dV/dT is given by ζ/eT and the thermoelectric force between two junctions at temperatures T_1 and T_2 is

$$V = \frac{k}{e}\int_{T_1}^{T_2} \log\left(\frac{nh^3}{2(2\pi mkT)^{3/2}}\right) dT. \qquad \ldots\ldots(3.17)$$

The thermoelectric power makes the hot junction positive in an n-type semiconductor, and negative in a p-type semiconductor. Thus by exploring the surface of a semiconductor with a hot metal probe connected to the material via a galvanometer, p- and n-type regions may readily be distinguished by the sign of the thermoelectric e.m.f.

The thermoelectric power at semiconductor–metal contacts is very large, of the order of 10^{-4} volt/°c.

3.2. *Optical Properties*

Radiation falling on a semiconducting crystal may break the valence bonds linking neighbouring atoms and produce a hole and an excess electron. In order that this process may occur the energy of a quantum of the radiation must be equal to or greater than the energy ϵ_i necessary to produce a hole–electron pair. This energy corresponds to a wavelength of hc/ϵ_i. ϵ_i may be determined from the temperature variation of the number of carriers in the intrinsic region (2.9) as determined by resistivity and Hall coefficient measurements. For silicon and

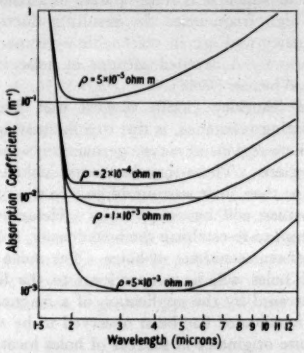


Fig. 8. Absorptivity of germanium (*after* Fan and Becker).

germanium (see table 1) the corresponding wavelength is 1·1 micron and 1·7 micron respectively. A sharp increase in absorption is therefore to be expected as the wavelength of the incident radiation is reduced through this critical value. The measured absorption in germanium as a function of impurity content and wavelength is shown in fig. 8 (Fan and Becker 1951) and similar curves are found for silicon. Beyond this sharp absorption edge silicon and germanium are very transparent to infra-red radiation. The absorption increases again towards long wavelengths and is larger for samples of lower resistivity, as would be expected if the absorption were due to the free carriers. The actual values, however, are much greater than is predicted by the theory of absorption by free carriers. Fan and Becker suggest that localized energy states due to impurities are important for this range of the absorption spectrum.

From the shift in the absorption edge with temperature, the variation of the energy gap ϵ_i with temperature may be determined (Fan 1950) and is found to be $-3{\cdot}5\times10^{-4}$ ev/°c and -4×10^{-4} ev/°c for silicon and germanium respectively.

The quantum efficiency of the light absorption process has been studied by Goucher (1950) by measuring the change in conduction of an n-type germanium filament due to illumination, which should be proportional to the total hole density. The number of holes present is equal to the product of the rate of generation and the lifetime, whilst the rate of generation is equal to the quantum efficiency multiplied by the rate of photon absorption. Goucher concludes that over a range of 1·0–1·8 micron, each photon absorbed produces one hole electron pair.

Emitter contacts can therefore be replaced by a light source as a source of injected holes, and use is made of this property in various photo devices, for instance the phototransistor (Shive 1949 b).

When an electron and a hole recombine, energy ϵ_i is released. As discussed in § 3.1.2 this energy may be emitted as a quantum of radiation or it may be exchanged with the thermal vibrations of the crystal lattice. Light emission due to the recombination of carriers has been observed (Lehovec, Accardo and Jamgochian 1951) in silicon carbide crystals in which the energy gap, about 2·5 electron volts, corresponds to radiation of wavelength of the order of 5 000 Å. Emission of radiation of wavelength about 1·7 micron has also been observed in n-type germanium carrying a current of holes. The quantum efficiency for this process is very small since most of the carriers recombine by exchanging energy with the lattice vibrations.

3.3. *Nucleon-irradiated Semiconductors*

Changes in conductivity and even of conductivity type can be produced in semiconductors by the passage of nucleons (deuterons, protons, neutrons or alpha-particles) and some of them even by the passage of electrons through the material (Lark-Horovitz 1951). The various types of phenomena responsible are:

(*a*) Transmutation. Exposure of silicon to thermal neutrons leads to the production of phosphorus atoms whilst fast neutrons produce aluminium atoms, the former acting as donors and the latter as acceptors. Deuterons and neutrons can give rise to many possible transmutations in germanium. If single crystal germanium is used then, provided that all disordering effects due to fast neutrons are healed out, an 'ideal' semiconductor may be produced, since the newly introduced atoms are definitely located at the lattice sites. The final conductivity will be due to the balance between carriers released from originally present impurities of either sign and from the impurity atoms introduced by transmutation. In germanium, slow neutrons produce more p-type impurities (gallium) than n-type impurities (arsenic) and lead to p-type material, the number of excess p-type impurities produced being 2% of the total neutron flux.

Comparing the data for integrated flux with Hall effect measurements after irradiation and heat treatment, it can be shown that each impurity centre releases one carrier. By careful control of the amount of irradiation it is possible to produce material which, apart from the increased impurity scattering, has the properties of an intrinsic semiconductor, because of a balance between the two types of impurities.

(*b*) Lattice defects produced by fast neutrons and charged particles. The irradiation of germanium with fast particles leads to some interesting effects (Brattain and Pearson 1950). Whilst very impure p-type specimens show an increase of resistivity, purer samples become better conductors after irradiation; n-type germanium is converted to the p-type form by bombardment, the resistance passing through a maximum as the original conduction electrons are neutralized, and then decreasing again when the material is converted to p-type (fig. 9).

Heat treatment at 450°C restores the original condition of the semiconductor, indicating that the effects are due to the production of lattice defects which are healed by the heat treatment. It is believed that both interstitial atoms (donors) and vacant lattice sites (acceptors) are produced by bombardment. The energy levels corresponding to both types of defect are below the intrinsic position of the Fermi level ψ ; thus the donors are rendered ineffective, and the p-type tendency results from the vacancy acceptors, which however have activation energies larger than those observed with chemically introduced acceptors.

(*c*) Electron bombardment. Experiments with electrons of energies from 0·7 MeV produce the same kind of permanent effects in germanium as heavy particles: changes from n-type to p-type. From the lower limit of energy necessary to produce this effect one can calculate that the minimum energy necessary to remove a germanium atom completely from its lattice site is about 30 eV.

Electrons of lower energies produce conductivity pulses and a decrease in resistance up to a factor of about 10, but these conductivity pulses do not last longer than about 10 microseconds. Moore and Herman (1951) have reported changes in forward and reverse currents in germanium point contact rectifiers resulting from bombardment by electron beams with energy of several kev.

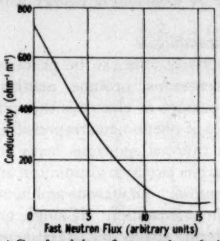


Fig. 9. Conductivity of germanium (originally n-type) after bombardment with fast neutrons (*after* Lark–Horovitz).

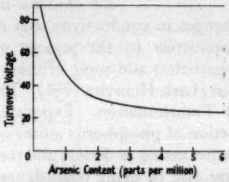


Fig. 10. Variation of turnover voltage of germanium point contact rectifiers with arsenic content of the germanium (*after* Douglas and James 1951).

(*d*) Positive ions. Ohl (1952) has described pronounced changes in the electrical properties of silicon produced by bombardment with positive ions. The effect is to increase the ratio of forward and reverse currents in silicon point contact rectifiers from a few hundred to more than ten thousand, and depends to some extent upon the type of ion gas used, helium being a preferred medium.

3.4. *Preparation of Semiconductors*

Requirements for semiconductor crystal devices have led to a demand for semiconducting materials with known and controllable electrical properties and have set new standards of purity and homogeneity. Indeed there is some difficulty in developing physical or chemical methods of determining the impurity content of semiconductors which are as sensitive as the electrical properties of the materials. The relation between resistivity and number of carriers in n-type germanium is shown in fig. 6. For the crystal diodes, triodes, etc. to be discussed in §5, germanium of resistivity about 0·1 ohm m is required, corresponding to an impurity concentration of about $5 \times 10^{20}/m^3$ i.e. one part in 10^8. The variation of turnover voltage of germanium point contact rectifiers with impurity content

is shown in fig. 10 and, in general, the higher the resistivity the more suitable the material for most applications.

Not only must the impurity content be small and carefully controlled, but the impurities must be uniformly distributed throughout the material. If this is not the case then, as explained in §2.3, the height of the Fermi level above the bottom of the conduction band will vary. But, since for equilibrium the Fermi level must be constant throughout the material, this means that the energy bands will be distorted and potential barriers are created which impede the flow of electrons and holes. Grain boundaries are undesirable for similar reasons. Barriers of this kind may act as traps for holes or electrons and reduce the lifetime and mobility of the carriers. Lattice defects and irregularities may act as acceptor centres and produce unwanted carriers as well as increased scattering. Most types of measurement—carrier lifetime, mobility, etc.—depend critically upon the degree of perfection of the crystal lattice, not only in the interior but also on the surface. Thus it is essential to remove any trace of a worked layer from the surface of germanium before making contact to it if good electrical performance is desired, and many etching treatments have been developed (Haynes and Shockley 1951). The important part played by the surface in the recombination of holes and electrons has already been mentioned in §3.1.2 and lifetimes are strongly dependent upon surface treatment.

Ingots of silicon or germanium prepared from very carefully purified material (Theuerer and Scaff 1951) in vacuum or in an inert atmosphere and using carefully cleaned crucibles etc. may, if carefully cooled, contain only a small number of crystallites. Segregation of impurities, due to different solubilities of the impurity atoms in the solid and liquid phase, can lead to an uneven distribution of carriers in the material but can also be used, by repeated meltings, as a means of purification. But semiconducting material of known and reproducible properties is probably best produced by the growth of single crystals from the melt under strictly controlled conditions. It is possible in this way to produce very pure uniform crystals, but the process is one of some difficulty.

If n-type germanium is heated to about 850°C and then quenched, p-type germanium is produced. If this same material is then heated for a number of hours at 450°C the material is n-type again after cooling slowly. These effects are attributed to the creation and subsequent healing of lattice vacancies, and it is suggested (Fuller, Theuerer and van Roosbroeck 1952) that the effect consists of the diffusion, from the surface inwards, of lattice vacancies which act as acceptors and cause the material to change to p-type. It is found by experiment (Fuller, Theuerer and van Roosbroeck 1952) that the concentration of lattice vacancies introduced into germanium by quenching from a temperature T°K is given approximately by

$$N_a = B\exp(-G/kT), \qquad \ldots\ldots(3.18)$$

where G is about 1·3 electron volts, and B is about 10^{28} m^{-3}. Shockley has pointed out that, whatever the nature of the acceptors introduced, B should be equal to the concentration of germanium atoms ($4{\cdot}52 \times 10^{28}$ m^{-3}).

The changes in carrier concentration and impurity type produced by nucleon bombardment have already been considered. Similar changes may be produced by allowing suitable impurities to diffuse into the body of the material (Fuller 1952, McAfee, Shockley and Sparks 1952). Recently developed semiconductor crystal devices depend on the properties of the transition region between n- and

p-type semiconductor material (p–n junction rectifiers, n–p–n and p–n–p junction transistors) and much effort has gone into the development of methods for making these junctions. Techniques based on the diffusion of impurities (Hall and Dunlap 1950) and on nucleon bombardment (Hall *et al.* 1951) have been described but, in principle, any of the methods described above for altering the conductivity type may be used.

§4. METAL–SEMICONDUCTOR CONTACTS AND p–n JUNCTIONS

4.1. *Historical*

Rectification phenomena with solid conductors have been known for many years, and practical use of these effects was made in the 'cats-whisker' crystal detectors of the early twentieth century, and commercial copper oxide and selenium rectifiers, long before a satisfactory theory was formulated.

It was confirmed that rectification was a contact rather than a bulk effect by experiments such as those of Schottky, Störmer and Waibel (1931). These workers demonstrated that almost all the voltage drop across a copper oxide rectifier biased in the reverse, or blocking, direction occurred in a very thin layer of oxide adjacent to the copper. Wilson (1932) showed that the contact between a metal and a semiconductor would show rectifying action if the two materials were separated by a thin potential barrier which could be penetrated by wave-mechanical tunnel effect. A serious defect of this theory was that a small change in the barrier thickness should affect the properties to an extent far greater than is observed in practice. An even more serious objection is that the direction of rectification given by the theory is opposite to that observed in practice.

More generally accepted theories of rectification also postulate potential barriers, but assume that these are too thick for tunnel effect penetration, so that electrons have to pass over the barriers. If tunnel effect penetration enters as a subsidiary of the principal rectifying process, the efficiency of rectification will be reduced as a result of this opposing tendency (Dilworth 1948).

4.2. *Metal–Semiconductor Contacts*

The first satisfactory theories of rectification at a metal–semiconductor contact were those of Mott (1939) and Schottky (1939). Both theories are, strictly speaking, applicable only for contacts involving semiconductors in which the mean free path of carriers is small, such as selenium or copper oxide, and have rather less validity when semiconductors such as silicon and germanium are involved. Nevertheless, they demonstrate the essential features of the rectification process, and will be considered below in some detail. These theories apply equally well for rectification between a metal and either an n-type or p-type semiconductor. In the p-type case, current passes most easily when the semiconductor has a positive bias, whilst an n-type semiconductor (as considered below) requires a negative bias to promote current flow.

The Mott and Schottky theories are based on the assumption that a potential barrier is set up between the metal and the semiconductor as a result of a difference in thermionic work functions. Electrons will tend to flow from the n-type semiconductor into the metal. This results in a positive charge on the semiconductor which depresses the electronic levels until dynamical equilibrium is reached, with the emission currents between the two solids equal and opposite.

This occurs when the Fermi level in the semiconductor is equal to that in the metal, as shown in figs. 11 (*a*) and 11 (*b*). In a semiconductor with a small impurity concentration, a small change in electron content is accompanied by a large shift in the Fermi level. This contrasts markedly with the highly degenerate metal, where the density of available levels at the Fermi energy is so large that the addition of electrons from the semiconductor affects the Fermi level to an entirely negligible amount.

When an external voltage V is applied, the Fermi levels at the two ends of the rectifier differ by eV, and a resultant current flows as a consequence of the disturbance of equilibrium. Once again, the Fermi energy of the metal is unaltered, and the adjustment of levels has to occur in the semiconductor, most of the potential drop appearing across the high resistance depletion layer, as shown in figs. 11 (*c*) and 11 (*d*). The potential by which the conduction levels are depressed in the absence of an applied voltage is V_d, termed the diffusion potential. When an applied voltage V makes the semiconductor negative with respect to the metal, the barrier height is reduced to $e(V_d - V)$, and the passage of electrons up the barrier and into the metal is greatly facilitated. This is the forward (easy current) direction of the rectifier. The forward current is slightly offset by a small flow of electrons from the metal over the barrier of height $\Phi_m - \chi$.

An applied voltage in the reverse direction depresses the levels in the semiconductor, and reduces the flow of electrons to the metal to negligible proportions. The resultant net current is thus just that from the metal into the semiconductor, still determined by the barrier height $\Phi_m - \chi$. In a good rectifier, this current should be as small as possible.

The Mott and Schottky theories both assume that carriers diffuse across the barrier layer, suffering many collisions with lattice vibrations in the process. Thus the current through the rectifier should be given by the sum of two terms, representing the effect of the electric field in the barrier layer, and of diffusion:

$$i_n = e\mu_n nE + eD_n(dn/dx). \qquad \ldots\ldots(4.1)$$

In this equation E is the electric field, and D_n the diffusion constant for electrons, which is given by the Einstein relationship (3.10). The applied voltage is $1/e$ times the difference in the quasi-Fermi level between the ends of the barrier layer, and the relationship between current and voltage can be obtained (e.g. Henisch 1949) by integrating (4.1). The result is

$$i_n = A\exp(-eV_d/kT)[\exp(eV/kT) - 1]. \qquad \ldots\ldots(4.2)$$

The quantity A depends on the properties of the semiconductor, and on the assumptions made about the distribution of impurity centres in the barrier layer. Thus Mott assumed that this layer is denuded of impurities, so that the field E is constant throughout the barrier, and in this case A is proportional to $V_d - V$.

Schottky assumed that the impurity concentration in the barrier is the same as in the bulk semiconductor, and the ionized centres in the depletion layer give a space charge eN_d, so that E varies with the distance from the contact. In this case $A \propto (V_d - V)^{1/2}$.

The actual shape of the potential barrier makes little difference to the general form of the current–voltage characteristic, which is in reasonably good agreement with experiment for rectifiers of the copper oxide and selenium type. Closer

agreement can be obtained when the reverse current is corrected for the Schottky effect of work function reduction in high fields, which results in an increased reverse current at high voltages.

For semiconductors such as silicon and germanium, in which the carrier mobilities are large, the mean free path may be comparable with or larger than the width of the barrier layer. Thus the flow of electrons across the barrier cannot strictly be described by the diffusion equation (4.1). Bethe (1942) considered the limiting case in which the barrier width is very small compared with the mean free path of carriers, so that the current from the semiconductor depends on the height, but not on the shape of the potential barrier. The current density for an applied voltage V on this 'diode' theory still has the general form of (4.2), and is given by

$$i_n = ne(kT/2\pi m)^{1/2}\exp(-\epsilon_b/kT)[\exp(eV/kT)-1]$$
$$= \tfrac{1}{4}ne\bar{c}\exp(-\epsilon_b/kT)[\exp(eV/kT)-1]. \quad \ldots\ldots(4.3)$$

In this equation, n is the free electron concentration in the bulk of the semiconductor, and $\bar{c}$ is the mean thermal velocity of an electron in a Maxwellian gas. The similarity of this equation to that for the rate of flow of particles through an aperture on the kinetic theory of gases is a natural result of the similarity of the problems.

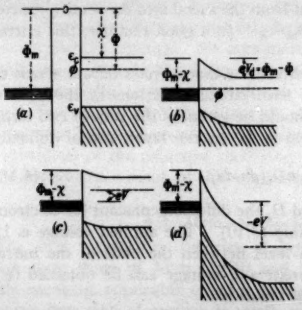


Fig. 11. The form of potential barrier set up at the contact between a metal of work function Φ_m and a semiconductor containing n-type impurities (not shown), with a smaller work function Φ, according to Schottky. (*a*) Before contact. (*b*) Equilibrium after contact for no applied voltage. (*c*) Voltage V applied in forward (easy current) direction. (*d*) Voltage V in reverse (blocking) direction.

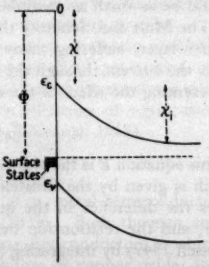


Fig. 12. The form of barrier set up at a semiconductor surface possessing a large density of surface states.

The diode theory as initially propounded still assumed that the barrier arose from a difference of work functions between metal and semiconductor, whereas Meyerhof (1947) found that V_d for silicon rectifiers was virtually independent of the metal used as a contact. As mentioned in §2.4, Bardeen (1947) has shown that the difficulty can be resolved by assuming a large density of additional levels at the semiconductor surface. In an n-type case, as shown in fig. 12, the existence

of low energy surface states will tend to attract electrons from the bulk semiconductor to these states, leaving a positive space charge layer just inside the semiconductor. When no metal contact is present, equilibrium is reached when the Fermi level in the interior equals that in the surface states, and the dipole layer formed by the negatively charged surface and the positive space charge layer has lowered the electronic levels in the interior by the amount χ_1. The amount of this depression depends on the number of surface states and their distribution in energy.

One of the most interesting features of this system lies in its behaviour when a metal is brought up to the semiconductor surface. Bardeen has shown that if the number of surface states is large, then χ_1 is not materially affected when a metal of arbitrary work function approaches to within a few ångströms, since the difference in work function can be compensated by changes in the surface charge instead of a space charge. This sort of spacing is likely to correspond with light mechanical contact in practical cases, since the two surfaces are separated normally by a thin film of oxide, moisture and dust. The partly filled surface states will have a metallic character if their density is large, and electron exchange between surface states and the metal will occur freely across the very thin insulating gap. The rectifying action, of either diffusion or diode type, occurs between the bulk semiconductor and its own surface levels. Thus the performance of a rectifier with this type of mechanism depends greatly on the type of surface treatment, but only in a minor degree on the nature of the metal contact used. This is found to be the case for many point contact rectifiers using germanium and silicon.

Rectifier theory has so far in this section been discussed on a one-current basis, that is, that hole conduction is negligible for rectification by an n-type semiconductor, and vice versa. This is not necessarily true, and has been shown to be very far from the truth in some practical cases. If the Fermi level near the surface lies nearer to the valence band than to the conduction band, a thin surface layer will have p-type properties, since the free hole concentration there is larger than the free electron concentration (Brattain and Bardeen 1948). When the influence of the valence band is considered in an n-type case, it is immediately obvious that the diffusion potential can never be greater than ϵ_1/e, since an attempt by either surface states or a metal contact to promote a larger value could be vitiated by removing electrons from a very small proportion of the large number of states at the top of the valence band.

When a metal contact is biased positively with respect to the semiconductor, the reduced electric field in the barrier will tend to encourage not only the diffusion of electrons out of the semiconductor, but also of holes away from the surface, into the truly n-type material, where they recombine with electrons (Haynes and Shockley 1949, 1951). It has been shown experimentally (Shockley 1951 a) that almost all the forward current of a germanium point contact rectifier consists of holes when n-type germanium is used, whilst electron current accounts for 60% of the forward current of p-type germanium rectifiers. The quantitative aspects of carrier injection, and its effect on rectification, can be more easily dealt with for the case of a planar p–n junction in which there is a mechanically continuous transition from p- to n-type material than for the theoretical models we have considered so far. The theory of the p–n junction will accordingly be considered next.

4.3. *p–n Junctions*

Much of the present widespread interest in p–n junction rectifiers and transistors has been aroused by the work started at Bell Telephone Laboratories by Shockley and his associates. This work has been based largely on considerations of rectification at a junction between n- and p-type forms of the same semiconductor (Shockley 1949), a problem which forms the main subject matter of this section.

The first theory of rectification at a p–n contact was that of Davydov (1938), who suggested that all practical rectifiers might be of the p–n type. Thus he regarded the copper oxide rectifier as a junction between copper rich (n-type) and copper deficient (p-type) cuprous oxide. In common with subsequent theories, Davydov concluded that current should pass most easily when a contact is so biased that the majority carriers on each side are encouraged to flow towards the interface, and he assumed that after traversing the contact, they were absorbed by recombination processes in the bulk material.

Gubanov (1950) has considered the behaviour of an abrupt transition from an n-type semiconductor to any p-type semiconductor. This theory, which allows for the effect of a thin film of adsorbed gas between the two solids, is based on the assumption that most current carriers are reflected from the interface, but that a net current results from recombination of electrons and holes at this interface. Injection of carriers from one semiconductor to the other, with subsequent volume recombination, is considered to be of very minor importance on this theory.

Shockley (1949) has considered very different conditions, and has concentrated on a study of the p–n junction transition within a single semiconductor sample. A very simple case to consider is that where donor and acceptor concentrations vary linearly with distance over a certain range, and are constant elsewhere, as shown in fig. 13 (*a*). The transition rate, a, across the junction is given by

$$N_d - N_a = ax. \qquad \ldots\ldots(4.4)$$

When no external voltage is applied to the contact, dynamic equilibrium is reached when the Fermi energy is constant throughout the solid. In order to satisfy this condition, holes and electrons are withdrawn from the central region, and a space charge dipole layer is set up which depresses the permitted levels on the n-type side with respect to those in the p-type material, as shown in fig. 13 (*b*). The shape of the variation in band energies with distance depends on whether the space charge of unneutralized impurity centres is important. Shockley derives expressions for the electrostatic potential as a function of position for cases in which the impurity gradient is either large (space charge important) or very small (space charge unimportant).

The net current of both electrons and holes separately across the junction is zero in the situation of dynamic equilibrium at zero bias. Thus considering the electron flow, some free electrons are continuously created in the p-region by thermal excitation, and those produced near the junction will be drawn down the potential gradient into the n-region. This is exactly compensated by a diffusion of electrons up the potential hill into the p-region, where they recombine with free holes. Similar considerations apply to the balance of hole flow across the junction.

This balance of flow is no longer found when a voltage is applied to the junction, since the diffusion currents are either increased or decreased, depending on whether the height of the potential barrier is decreased or increased, whereas the reverse currents, arising from carriers produced by thermal ionization, are virtually unchanged. For voltages of more than a few kT/e applied in the reverse direction (n-type semiconductor made positive), the diffusion currents become negligibly small, and the saturation reverse currents I_{ps} and I_{ns} of holes and electrons respectively are found.

In the forward direction of the rectifier, considerable diffusion currents occur, and the injected carriers may travel for some distance before they are captured. In consequence, the concentrations of both holes and electrons in the junction are larger than the thermal equilibrium values, and should be described in terms not of the Fermi level ϕ, but of the quasi-Fermi levels ϕ_n and ϕ_p, which coalesce at ϕ for regions far from the junction. This separation of the Fermi level into two components is shown in fig. 14(*b*), where a rather idealized view of a junction is taken. The final values of both ϕ and ψ towards the outer end of the p-type region are both shifted downwards by eV when this region is biased V volts positive.

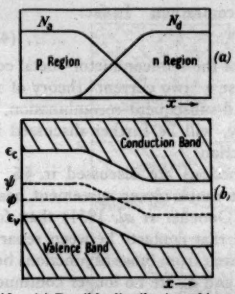


Fig. 13. (*a*) Possible distribution of impurities in a p–n junction. (*b*) Corresponding variation of electrostatic energy with distance across the junction for zero applied bias.

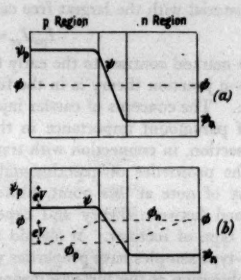


Fig. 14. The energy ψ corresponding to the midpoint of the bands, and its variation with respect to the Fermi level through a p–n junction. (*a*) At zero bias. (*b*) With a voltage V applied in the forward direction.

The currents flowing at any point depend on the slopes of the quasi-Fermi levels. From (3.11)

$$i_p = p\mu_p \frac{\partial \phi_p}{\partial x}. \qquad \ldots\ldots(4.5)$$

Similarly the current density due to electrons is

$$i_n = n\mu_n \frac{\partial \phi_n}{\partial x}. \qquad \ldots\ldots(4.6)$$

From these equations it can be seen that the major part of the change in ϕ_n from one side of the junction to the other occurs in the p-region. Similarly ϕ_p changes most rapidly in the n-region, and $\phi_n - \phi_p$ at the middle of the junction may be only a little less than eV. This is an excellent state of affairs, since $V - (\phi_n - \phi_p)/e$ is equivalent to a voltage drop across an undesirable ohmic

resistance R_1 in series with the rectifier. By considering the rates of flow of electrons and holes for a 'good' rectifier, where this ohmic voltage drop is small, Shockley deduces a relationship between voltage and current of the form

$$I = I_s[\exp(eV/kT) - 1]. \qquad \ldots\ldots(4.7)$$

This is identical in form with (4.3) for a metal–semiconductor contact on the diode theory. I_s is the saturation reverse current for voltages from a few kT/e up to the breakdown voltage and can be expressed as the sum of electron and hole currents,

$$I_s = I_{ns} + I_{ps}. \qquad \ldots\ldots(4.8)$$

Whatever the applied voltage, the ratio of electron current to hole current should always be I_{ns}/I_{ps}.

The value of I_{ns} depends only on the properties of the p-type material, since it can be written in the form

$$I_{ns} = kT\mu_n n_i^2/pL_n. \qquad \ldots\ldots(4.9)$$

In this expression, p is the normal hole concentration in the p-material, and the diffusion length $L_n = (D_n\tau_n)^{1/2}$ is an average distance covered in this material by an electron before capture. A similar expression holds for I_{ps}, and the not unnatural result appears that the largest contribution to the current comes from the material with the largest free carrier concentration. In fact

$$I_{ns}/I_{ps} = (\sigma_n/\sigma_p)^{1/2}. \qquad \ldots\ldots(4.10)$$

In marked contrast to the early theories of the semiconductor–metal contact, this p–n junction theory is in the fullest sense a 'two current' theory of rectification. The concepts of carrier injection and subsequent recombination, which are of paramount importance in this theory, will be further discussed in the next section, in connection with transistor action.

The properties of experimental p–n junctions are discussed in §5. It is worthy of note at this point, however, that much closer agreement has been obtained between theory and experiment (Goucher *et al.* 1951) than for any other type of rectifier. It should be noted that contacts between separate p- and n-type samples have properties very different from those of junctions because recombination at the surfaces means that ϕ_n and ϕ_p are no longer continuous.

4.4. *Transistor Action*

The discovery of transistor action has increased enormously the scope of semiconductor applications. Some discussion of current views on the basis of transistor action would be appropriate before a consideration of the present state of development of these devices, and is given in this section.

In its first and simplest form, the Type-A transistor (Bardeen and Brattain 1948), the transistor comprised a block of n-type germanium to which three connections were made. One of these, on the base of the block, was a large area electrode, deliberately made non-rectifying, whilst the others were closely spaced whisker point contacts, which separately showed normal rectifying properties.

One of these, usually termed the collector, was negatively biased, so that a small current flowed. It was found that when the other point contact, the emitter, was positively biased, producing a current I_e, then the collector current I_c was augmented by an amount of the same order of magnitude as I_e. Variations of I_e were accompanied by similar variations of current in the collector circuit.

As the impedance of a negatively biased rectifier is very much greater than the forward resistance, the power output of the collector can be many times greater than the power input to the emitter.

Considerable evidence supports the view that transistor action occurs when (in an n-type case) holes are injected from the positively biased emitter, and drift towards the collector, where the strong electric field tends to draw them into the negatively biased point contact.

As mentioned in § 4.2, an appreciable fraction γ of the current carried by a contact biased in the forward direction consists of minority carriers 'injected' from the metal. For contacts on n-type germanium, the emitter current consists almost entirely of holes.

The electric field in the semiconductor encourages the drift of these holes towards the collector, but only a fraction β will arrive, the remainder having been absorbed by recombination with free electrons. Thus a hole current $\beta\gamma I_e$ arrives at the collector. The value of β depends on the distance between contacts and on the hole lifetime τ_p. In favourable circumstances, $\beta\gamma$ may be not far short of unity. The current gain can however be larger than $\beta\gamma$ since 'current multiplication' effects can be produced at the collector contact.

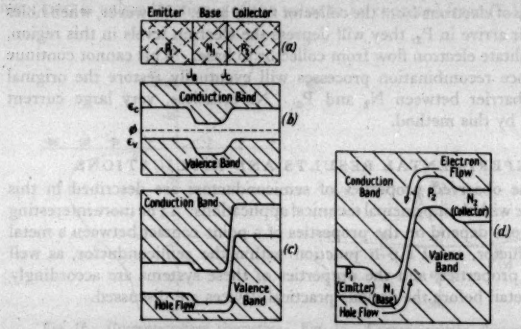


Fig. 15. (*a*) Arrangement of zones in a p–n junction transistor. (*b*) Variation of band energies in the absence of external bias. (*c*) With normal operating bias applied. (*d*) p–n–p transistor with hook collector.

Shockley (1951 a) has discussed a way in which such an effect can occur, by assuming that the presence of an appreciable density of holes near the collector promotes the emission of electrons from the metal point, in order to restore electrical neutrality. These electrons travel to the base electrode. Thus the collector current is enhanced not only by the flow of holes from the emitter, but also by the flow of neutralizing electrons from the collector. Values of current gain up to $1+\mu_n/\mu_p$ can be explained in this way, and larger values observed in practice can be attributed to the influence of hole traps (Sittner 1952) or of a p–n hook (Shockley 1951 a) as discussed in § 5.

The point contact transistor as a whole is not amenable to precise calculation of properties. The theoretical problem is simplified if the geometry is made one-dimensional. In the junction transistor, the point contacts are replaced

by area p–n junctions on either side of a thin slice of base material. This is illustrated in fig. 15 (*a*) for a case in which the central zone is n-type, placed between broader p-type regions.

In the absence of an applied bias, the band energies will vary in such a way that the Fermi energy is constant throughout, as shown in fig. 15 (*b*).

In operation, the collector (region P_2) is made strongly negative, so that the junction J_c is in a state of high resistance. At the same time, a small positive bias is applied to the emitter P_1 which causes an appreciable current to flow across the junction J_e. If P_1 has a greater free carrier concentration than the base N_1 almost all the emitter current will consist of holes flowing into the base. These will very rapidly pass across J_c and into the collector, suffering only a very slight attenuation due to recombination in the base if this layer is made sufficiently thin. From this model, it is obvious that the changes in collector current cannot be larger than those in emitter current.

Junction transistors can be produced with a current gain greater than unity by incorporation of a p–n hook collector. For the p–n–p construction of fig. 15, this is accomplished by adding a further n-type section N_2 after P_2 (fig. 15 (*d*)). The existence of a contact potential difference between N_2 and P_2 will normally prevent the flow of electrons from the collector to the base. However, when holes from the emitter arrive in P_2, they will depress the electron levels in this region, and greatly facilitate electron flow from collector to base. This cannot continue indefinitely, since recombination processes will eventually restore the original height of the barrier between N_2 and P_2. Nevertheless, very large current gain is possible by this method.

§5. EXPERIMENTAL RESULTS AND APPLICATIONS

Some of the observed properties of semiconductors are described in this section, together with their principal technical applications. The more interesting circuit applications depend on the properties of a point contact between a metal and a semiconductor, or of a p–n junction within the semiconductor, as well as on the bulk properties, and the properties of these systems are accordingly considered in detail before the various practical devices are discussed.

5.1. *Properties of Point Contacts*

Rectification at a metal point contact on silicon has been known for some time, and the silicon microwave mixer crystal received intensive development during the last war. This work has been reviewed elsewhere, however (Torrey and Whitmer 1948), and the emphasis here will be on more recent work on germanium.

The relation between voltage V and current I at a contact on n-type germanium may be divided into two main regions according to the direction of current flow (fig. 16). The low-resistance forward region and high-resistance inverse region will be considered separately.

5.1.1. *Forward Characteristic*

No matter which of the mechanisms considered in §4.2 is postulated, the predicted relation between current and voltage is of the form, due originally to Wagner (1931),

$$I = I_s[\exp(qV/kT) - 1]. \qquad \ldots\ldots(5.1)$$

The voltage V' measured between the metal point and a low-resistance area contact to the semiconductor (base contact) must be corrected for a voltage drop in the unavoidable series resistances. The base contact itself may introduce a small series resistance of this kind. In silicon the base contact has been found to be almost non-rectifying, with a resistance for unit area of about 10^{-6} ohm m^2 (Putley 1947), whilst in germanium it has been suggested that the transition from the metal of a soldered contact to the semiconductor is of a continuous nature, and introduces little extra resistance (Pantchechnikoff 1950). Although the base contact is never in fact completely non-rectifying, the voltage drop across it may usually be neglected in comparison with that across the spreading resistance R_s, which arises from the constricted current flow in the semiconductor near the rectifying contact. The barrier voltage V is thus given by

$$V = V' - IR_s. \qquad \ldots\ldots(5.2)$$

In practice it is found that values for I_s and R_s can be chosen which satisfy a straight line relation over part of the range when $\log(I+I_s)$ is plotted against V, but that no one pair of values can be assumed which fits the observations over the whole of the forward characteristic (Yearian 1950); moreover, the slope of the linear part is only a fraction η of the value to be expected from (5.1). Since this

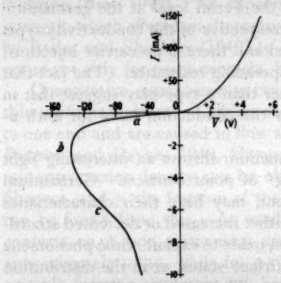


Fig. 16. Current–voltage characteristic of a contact between metal and n-type germanium (note change of scale). (*a*) Low-voltage region of inverse characteristic. (*b*) Turnover region. (*c*) Negative resistance region.

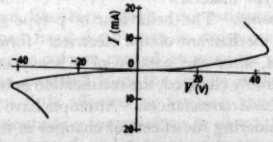


Fig. 17. Current–voltage characteristic of a contact between metal and p-type germanium (*after* Bray 1946).

quantity can be derived more accurately from the inverse characteristic, it will not be considered again until §5.2. If I_s is taken as constant, the value of R_s which satisfies (5.1) and (5.2) usually decreases progressively with increasing current, reaching values which are an order of magnitude less than that calculated (Torrey and Whitmer 1948, p. 427) from the radius of the contact and the bulk resistivity of the semiconductor. Explanations based on resistive heating of the contact may be discarded, because the effect occurs at a power dissipation too low to produce the necessary temperature rise. An attempt has been made to account for the effect by assuming that the resistivity decreases with increasing electric field, but the observed changes in resistivity of n-type germanium in high fields are

of the opposite sign to that required to explain the form of the forward characteristic (Shockley 1951 b), and it now seems certain that the correct explanation is in terms of the surface states discussed in §§ 2.4 and 4.2 (Brattain and Bardeen 1948). The p-type layer produced by these states can add a small surface conductance, although not enough to account for the observed values of R_s, but the hole injection accompanying forward current flow will, by increasing the bulk conductivity, add a conductance which increases with current, and which appears to be of the order of magnitude required to account for the observed characteristic. The existence of an analogous layer on p-type material is doubtful, but since injection occurs in both p- and n-type silicon (Haynes and Westphal 1952), and in some contacts on p-type germanium (Shockley, Pearson and Haynes 1949), explanations based on injection may apply to these materials also.

Contacts on p-type germanium, if they have received no special electrical treatment, have properties quite different from those of similar n-type contacts. The characteristic (fig. 17) often shows a nearly symmetrical high resistance (corresponding more nearly to the calculated value of R_s) for applied voltages of up to 20 v or more. At higher voltages the current generally increases more rapidly in what would be expected to be the forward direction (metal point negative), although characteristics with the opposite sign of rectification are not unknown—nor are nearly symmetrical curves (Dunlap and Hennelly 1948). This result is consistent with the view that the Fermi level at the germanium surface is near the top of the valence band irrespective of the conductivity type. In this case, only a very low barrier is formed and there is no carrier injection; thus the resistance of the contact is just the spreading resistance. The fact that p-type silicon generally forms a better rectifier than n-type may suggest that in silicon the Fermi level at the surface is near the conduction band for both n- and p-type material.

Forming. The behaviour of p-type germanium throws an interesting light on the mechanism of the electrical 'forming' of point contacts. Germanium contacts, after the passage of a heavy current, may have their characteristics considerably changed, the rectification being either increased or decreased according to the circumstances. Attempts have been made to explain these phenomena by considering the effects of changes in the surface states, or in the distribution of impurity centres and traps near the contact. There is some doubt as to the roles of the various possible mechanisms, since hypotheses which explain changes in the rectification characteristic are inadequate to account for the observed variations in the effect of minority carriers on the contact (see § 5.1.2) and vice versa. It is probable that with small currents changes occur only in the surface states, but that with higher currents other effects become more important.

Bardeen and Pfann (1950) have explained the rectification changes in germanium by assuming that the passage of current from metal to semiconductor (positive forming) can produce a permanent increase in $(\Phi - \chi)$ (fig. 4 (*d*)), whilst flow in the opposite direction (negative forming) can produce a permant decrease. These changes alter the barrier height, and so the rectifying properties of the contact.

The barrier height at an unformed contact on n-type germanium is large, so that the decrease in inverse current produced by positive forming should be small compared with the increase resulting from negative forming. At a contact on p-type material on the other hand, the barrier is initially low, so that the

inverse current should be increased by positive forming and greatly decreased by negative forming. In this last case the increased barrier height should also lead to greatly increased electron injection, and so to a decrease in R_s and an increase in forward current. With one exception, these are the changes observed when currents of the order of an ampere are passed for a fraction of a second.

Characteristics similar to those of n-type contacts can be obtained from p-type material in this way (Pfann and Scaff 1950), although the expected decrease in inverse current is not observed. This suggests that the barrier height may not be the only factor affecting the inverse current. That it is the semiconductor which is altered is shown by the fact that the metal point may be removed and replaced by another without materially changing the characteristic.

Carrier Injection. The recognition that part of the forward current at a rectifying barrier may be carried by minority carriers provides the basis for most of the recent developments in semiconductor devices. This concept was first proposed as an explanation of transistor action (§ 4.4) and of the observed variation of R_s with current. It was at first thought that the injected carriers were confined to a thin surface layer, but experiments in which holes are injected on one side of a thin septum of n-type germanium and detected on the other (Kock and Wallace 1949) show that they flow through the volume of the specimen. This is confirmed by measurements of the resistivity and Hall coefficient during injection (Bray 1949). An apparent dependence of resistivity on electric field strength, which offered an alternative explanation of some of the effects, was shown by Ryder and Shockley (1949) to be the result of carrier injection. This can occur not only for point contacts, but also at soldered and plated contacts.

Quantitative measurements can be made using a specimen in the form of a thin uniform filament. Carriers are injected into this from a contact placed near to one end and are caused to flow along it by an applied electric field (Shockley, Pearson and Haynes 1949, Haynes and Shockley 1951). An estimate of the minority carrier density can be obtained from the change in resistance of the filament, and γ, the fraction of the emitter current carried by minority carriers, can be found from this. As would be expected from the results on forming, contacts on n-type germanium are but little affected by forming and have γ approximately unity, while for p-type contacts γ increases up to about 0·6 with suitable forming treatment. γ has also been found to vary with temperature (Bray 1949).

When the density of minority carriers differs from the equilibrium value, some of them will tend to diffuse towards an isolated metal point which makes a rectifying contact with the semiconductor. In consequence, the potential of the metal point changes until the flow of carriers towards it is just balanced by injection. The potential thus acquired, the 'floating potential' V_f, forms a useful measure of the minority carrier density, but may be a disturbing influence in probe measurements of potential. The observed variation of V_f with carrier density and contact properties is in agreement with the theory given by Bardeen (1950).

Effect of Temperature. In that region of the forward characteristic where the current is limited mainly by R_s (V greater than a few tenths of a volt), the effects of temperature are small. The current for a given applied voltage rises slightly as the temperature increases above room temperature, although the bulk conductivity is still falling in this range. Since this is the region where injection

effects are important, the explanation may lie in changes in either γ or carrier lifetime. Where the barrier controls the current, the situation is more complicated. The fraction η appears to decrease with increasing temperature for unformed contacts and to increase for formed contacts (Yearian 1950). In both cases the current at constant voltage increases with an activation energy of about 0·3 ev.

Effects of Forward Overloads. The passage of large forward currents may produce changes in the properties of a contact. With currents of less than about 0·2 amp, only reversible changes due to heating appear. Between these currents and about 1 amp, permanent changes due to forming may appear, the inverse resistance at first rising and then falling sharply until, after a really severe overload, the contact may be almost non-rectifying. There appears to be a (thermal) time-constant associated with these changes, since effects which occur with current pulses of a few tens of milliseconds are often absent with microsecond pulses.

5.1.2. *Inverse Characteristic*

The inverse branch may be divided, as in fig. 16, into three parts; the low voltage region, where heating effects are negligible, the turnover region, where they become important, and the negative resistance region, where they are dominant.

Low Voltage Region. In the low voltage region, the relation between I and V may be compared with the diode theory of § 4.2. On this basis, one would expect the current to increase rapidly according to (5.1) to a saturation value I_s, at an inverse voltage of the order of $\boldsymbol{kT/e}$ (about 25 mv at room temperature). However, the saturation expected is never in fact observed, and the characteristic is often more nearly that of a high ohmic resistance. A straight line can often be drawn through the measured points near the origin, and the intercept on the current axis can be interpreted as a saturation current. From the slope of the (I, V) curve at very small voltages η can be obtained with more certainty than from forward current measurements, since injection effects are absent. Values of η obtained in this way rarely exceed 0·5, and the absence of an adequate explanation of this fact represents one of the major gaps in our understanding of contact phenomena.

Attempts have been made to explain both the anomaly in η and the absence of saturation by assuming that there exists on n-type germanium a p-type surface layer whose transverse conductance is large enough to carry an appreciable part of the current, and that the rectifying barrier occurs between this layer and the n-type interior of the semiconductor. The area of the barrier should then vary with applied voltage in a way which would lead to low values of η at small voltages, and to a nearly ohmic characteristic at high inverse voltages (Aigrain 1950). In practice, the conductance found on normally treated surfaces is insufficient to account for the observed characteristics.

An alternative explanation has been given in which the barrier height is assumed to fluctuate from point to point over the contact area (e.g. Johnson *et al.* 1950). Small patches where the barrier is low will behave as low resistance rectifiers with large spreading resistances in series. An approximation can be obtained to almost any characteristic by the choice of suitable distributions of barrier height and patch area.

Surface treatments have a marked effect on this part of the characteristic (Hogarth and Granville 1951). A freshly cleaved germanium surface usually exhibits a fairly high inverse resistance, but on one which has been mechanically worked (ground, polished or sandblasted) the inverse resistance may be so low as to make the contact almost non-rectifying. If the worked surface is etched the resistance rises to values usually a little larger than those observed on the cleaved surface. Many processes have been described, but in practice there seems to be little to choose between them, provided they remove the worked layer. It has also been shown that if the worked layer is recrystallized by heat treatment, properties similar to those of a cleaved or etched surface are obtained. The effect of material purity is often obscured by other effects, but in general the resistance is higher for contacts on purer material.

The low voltage inverse current is also sensitive to the concentration of minority carriers near the contact. The term collector is used to describe a contact in which this effect is utilized. When the voltage is less than about 3 v, minority carriers diffuse to the contact and increase the inverse current, while voltages of about 20 v set up a field of potential flow which attracts many more carriers to the contact and so produces a larger current change (Bardeen 1950). No results on the variation of these effects with voltage seem to have been published, but it is known that forming treatments which increase the (normal) inverse current improve the sensitivity of the collector to minority carriers, partly, perhaps, by increasing the volume from which carriers are extracted.

In the case of an n-type semiconductor where the normal electron concentration is n_0, and the reverse current I_0, it can be shown that the change in current due to an addition δp to the hole concentration should be

$$\delta I = I_0 \, \delta p / b n_0, \qquad \ldots\ldots (5.3)$$

where $b = \mu_n/\mu_p$. If $\delta I'$ is the change actually observed, the current multiplication factor α' is given by

$$\alpha' \equiv \delta I'/\delta I. \qquad \ldots\ldots (5.4)$$

Because early n-type transistor collector contacts had a multiplication factor of about 2·5, it was thought that the current change was brought about by a change in conductance of some part of the semiconductor. This seems unlikely, since the electron component would have to traverse the reverse-biased barrier at the contact. A possible explanation is that the space charge of the holes arriving at the barrier provokes, for example by field emission, the emission of an abnormally large number of electrons (Bardeen and Brattain 1949). These will set up a space charge which just balances that of the holes when the ratio of the currents is equal to the ratio of the mobilities, so that values of α' up to $b+1 = 3{\cdot}1$, can be explained, but not the higher values of over 20 which have been observed. Theories involving the production of secondaries can be excluded because they would be expected to indicate an exponential dependence of α' on collector voltage, in contrast to the nearly voltage-independent behaviour observed. A theory capable of explaining the observed values can be derived from the field emission theory (Shockley 1950 b) if a mechanism equivalent to a local increase in b is postulated. One such theory assumes that there is, near the barrier, a number of traps in which the holes spend a fraction θ of the time, so that the mobility ratio is increased to $b/(1-\theta)$. Arbitrarily large values of α' can thus be explained. This theory also accounts for the fact that large values of α' are found only for the smallest values of $\delta I'$, and that α' decreases to about $b+1$ for large

currents (Sittner 1952). This is the result to be expected if the density of traps is rather low, and they become saturated. The dependence of α' on $\delta I'$ has been derived for this model and is in fair agreement with experiment. Measurements of the effects of temperature on this dependence lead to a value of about 0·3 ev for the depth of the hole traps and about 10^{19} m^{-3} for their density. The latter decreases with rising temperature, suggesting that the traps are distributed in depth. The rough thermal treatment which the collector receives during forming is quite capable of introducing densities of this order.

A second theory involves a 'hook' structure like that discussed in § 4.4. No quantitative discussion has been given which would account for the variation of α' with current, but rough calculations indicate that on this model also the general shape of the curve should be similar to that observed in practice. It is thus difficult to decide between the two theories on evidence of this sort, a not unexpected result in view of the trap-like nature of the 'hook'. Direct evidence supporting the latter has, however, been obtained (Valdes 1952). Probe measurements indicate that heavy forming of a phosphor bronze contact produces a high resistance barrier at a radius of about 100 microns from the contact, and a region of increased conductivity within this radius. These effects disappear when a layer about 10 microns thick is etched from the surface, and they are interpreted as being due to an 'island' of p-type material. When the forming has been extremely heavy the p-type nature can be demonstrated by measurements of the thermoelectric effect. Neither of these methods is capable of detecting the thin n-type layer immediately under the contact which is an essential feature of the 'hook' collector. Its existence can be inferred from the fact that high values of α' are obtained only when the contact point is made from a metal, such as phosphor bronze, containing a donor impurity. It is thought that in this case both phosphorus and 'lattice defect' (§ 5.2) diffuse out from the contact during forming. The p-type lattice defect diffuses more rapidly and so forms the intermediate layer. It should be noted, however, that the methods described above are not sufficiently sensitive to determine whether or not such effects occur in normally formed contacts.

The inverse current contains a large excess noise component, whose origin is not understood. Most of the published work relates, not to the noise from a single contact under d.c. conditions, but to that arising from operation as a mixer or as the collector of a transistor.

The effect of temperature on the low voltage inverse current in n-type germanium is complex, both the 'saturation current' and the ohmic components discussed above increasing rapidly with temperature. The saturation current can be fitted by an expression of the form

$$I_s = A \exp(-\epsilon_b/kT), \qquad \ldots\ldots (5.5)$$

with ϵ_b, the apparent barrier height, equal to about 0·67 ev, compared with 0·76 ev for ϵ_i. From measurements of the quantity $(dI/dV)_{V=0}$, which should be proportional to I_s, a value for ϵ_b of about 0.60 ev is obtained. Similar measurements on p-type silicon for which $\epsilon_i = 1{\cdot}1$ ev give the curiously low value of $\epsilon_b = 0{\cdot}38$ ev (Billig and Ridout 1951).

The behaviour of the ohmic component is somewhat erratic and little information has been published on the subject. It is often subject to thermal

hysteresis, and depends on the method of preparation of the contact. The increase of this component often sets an upper limit to the temperature at which germanium rectifiers are useful, but it is interesting to note that silicon might be a better material for use at high temperatures since ϵ_i is larger. Until recently, silicon contacts have always had a much lower inverse resistance, but recent work on the positive ion bombardment of silicon surfaces gives hope of great improvement (Ohl 1952).

Since radiation of wavelength shorter than about 1·8 microns can produce hole-electron pairs in germanium, a collector contact will respond to suitable illumination (Shive 1949 b). The quantum efficiency may exceed unity at about 1·5 microns, owing to the fact that $\alpha'>1$ for some contacts, and drops off for shorter wavelengths, because the light is absorbed so close to the surface that the effective lifetime of the carriers is reduced. The sensitive area extends for a fraction of a millimetre around the contact; from this radius the diffusion length, L_p or L_n, can be found. Photovoltaic effects are also observed, owing to floating potential changes.

High Inverse Voltage Region. In the turnover region, the resemblance between the observed characteristics and those expected on any of the simple theories discussed in §4.2 tends to be obscured by thermal effects. The current begins to rise with increasing rapidity as the voltage is raised until, at the turnover voltage V_T, dV/dI becomes zero. A second part of the characteristic, on which the incremental resistance is negative, extends from this point towards lower voltages and higher currents (fig. 16). Both turnover and negative resistance regions may show hysteresis effects when first explored, and even after reproducible conditions have been obtained the current may increase slightly for a few milliseconds when a voltage is applied.

It is thought that these effects are due to the reduction of the resistance of the contact by resistive heating. This view is supported by measurements at various ambient temperatures of the power dissipated at turnover. The power decreases linearly with increasing temperature over quite a wide range, thus suggesting that the contact temperature at turnover is approximately constant. The magnitude of this temperature can be found by extrapolating to zero power, and the value so obtained lies in the region where intrinsic conduction is becoming important (Benzer 1949).

Characteristics taken with pulses shorter than about 10 microseconds duration, which is apparently less than the thermal time-constant, never show the negative resistance region, and have been reported to be nearly linear up to at least twice V_T (Bennett and Hunter 1951). The static characteristic can be derived from the short-pulse, or isothermal curves by a geometrical method. If the isothermal characteristics are assumed to be linear, with an activation energy for the conductance of about 1 ev, and the contact temperature is taken to be proportional to power, a curve is obtained which is certainly of the right shape. The quantitative agreement is poor, however, the measured static resistance sometimes being higher than the isothermal value, and the observed activation energy is much less than 1 ev.

The turnover voltage is sensitive to the same influences as the low-voltage resistance, and varies in the same general way. Its variation with the impurity content of the germanium has been investigated by Douglas and James (1951)

whose results show that V_T varies approximately as (arsenic content)$^{-1/3}$ (fig. 10). Increased temperature results in a decrease of V_T, for example from 50 v at 20° c to 20 v at 100° c. A significant increase in V_T can be obtained by positive forming.

The negative resistance region has not attracted great interest, since contact behaviour here is often erratic. Oscillations are sometimes found, which do not seem to be relaxation oscillations of the external circuit, but to be a property of the contact. If the current is increased to more than a few times the turnover value, irreversible forming effects take place.

5.1.3. *Signal Properties*

The small-signal ($|V| \leqslant kT/e$) properties of silicon and germanium contacts have been discussed by Torrey and Whitmer (1948) and recent progress has been largely technological. Silicon remains the better detector and mixer at high frequencies, although germanium is now receiving attention as a mixer for frequencies of about 10^3 Mc/s. Germanium contact rectifiers are finding wide application at higher signal levels ($V \gg kT/e$), both with sinusoidal and with pulse signals, and their behaviour under these conditions has been studied in some detail.

The results of measurements made with sinusoidal signals of a few volts amplitude are usually expressed as the 'rectification efficiency', an empirical quantity defined in terms of the d.c. output from a specified circuit in which the contact is worked as a rectifier under standard conditions. It is found that the decrease in 'efficiency', which always occurs at high frequencies, becomes serious at lower frequencies for contacts with high turnover voltages (Douglas and James 1951). This probably indicates only that those variables, such as the conductivity of the germanium, which affect V_T also affect the frequency response. This could be due either to an effect like the Meacham–Michaels effect (see below), which would reduce the inverse resistance at high frequencies, or to the finite transit time of injected holes. When this becomes comparable with the period of the applied voltage, the decrease in R_s is prevented, and the effective forward resistance is greater. Both of these effects would be more noticeable in low-conductivity material, but the time-constant associated with the first of these seems to be too large to account for the observed results. Where these effects can be avoided, and the peak signal voltage kept to rather less than V_T, germanium contacts form fairly accurate linear detectors.

The capacity of the contact itself is difficult to measure and of doubtful significance, since the true area is so uncertain. Measurements on commercial units indicate that this capacity is of the order of 1 pF and that it increases with forward voltage and decreases with inverse voltage, in qualitative agreement with Schottky's theory.

When measurements are made with voltage pulses of a few microseconds duration, several additional effects are found. The inverse characteristic becomes almost linear, as discussed above, and in some contacts the forward current does not at once rise to its static value. In other, apparently identical, contacts this latter effect does not occur, and its cause remains obscure.

Meacham and Michaels (1950) have observed that in certain contacts the application of a pulse of inverse voltage is accompanied by an abnormally large current if the contact has previously been passing a forward current of some tens of milliamperes. In those contacts which show the effect, a large inverse current

begins to flow about 0·05 microsecond after the application of the inverse voltage and dies away, with a variety of waveforms, to the static value within about 0·4 microsecond. The current may reach 100–200 mA compared with the static value of 1 mA. It is thought that this current represents the collection by the contact of previously injected holes which have not recombined. Currents of the observed magnitude could be accounted for by a hole density of $2{\cdot}5 \times 10^{20}$ m^{-3} in a volume of about one cubic millimetre. Other effects have been observed in which the stored holes apparently give rise to a floating potential.

5.2. *Preparation of p–n Junctions in Germanium*

Although natural p–n junctions are sometimes found in rather impure ingots, such impure material is unlikely to make good rectifiers since the diffusion lengths L_n and L_p will be too small. It is evident from the theory of § 4.3 that junctions in relatively pure material should form excellent rectifiers, and much work has been done on the preparation of such junctions. In some of the earliest attempts use was made of the changes which germanium undergoes when bombarded by slow neutrons or by fast nucleons (§ 3). In this method part of an n-type germanium specimen is irradiated, while the rest is kept screened, so that the exposed part becomes p-type, and a rectifying junction is formed. Alternatively, short-range particles such as α-particles from polonium can be used, one face of the specimen being exposed and becoming p-type to a depth depending on the range (Hall *et al.* 1951). However, neither method results in good junctions. Apart from the difficulty of controlling the transition rate, there is the fundamental difficulty that the lattice disorder introduced reduces L_n and so the rectification. Some of this disorder can be removed by annealing, but irradiation methods are now but rarely used.

In a second method of preparation the impurities in the melt are varied during solidification. If freezing takes place progressively, as is necessary for the production of homogeneous material, a junction forms in the solid ingot at a position corresponding to that of the liquid–solid interface when the net impurity type of the liquid changes. Teal, Sparks and Buehler (1951) describe an example of this method in which the impurity content of the melt is changed during the withdrawal of a seeded crystal, so that a single crystal containing a junction is obtained. Since crystals grown in this way are nearly perfect, lifetimes up to 600 microseconds are found, and good rectifiers are obtained.

In another method, impurities are diffused at high temperature through the solid germanium crystal (Hall and Dunlap 1950). The density N_1 of a diffusing impurity at a depth x after a time t should be given by

$$N_1 = Q(\pi D_1 t)^{-1/2} \exp(-x^2/4D_1 t), \qquad \ldots\ldots(5.6)$$

where D_1 is the diffusion coefficient of the impurity and Q its initial concentration in atoms per unit area at the surface $x = 0$. If this impurity is of the opposite type to one already present in uniform concentration N, the conductivity type will change with the sign of $N - N_1$, and a junction should be formed where $N = N_1$. That a surface layer of impurity does behave in this way is shown by measurements with radioactive tracers (Dunlap and Brown 1952), the results of which are in good agreement with results derived from the position of the junction as located by observing the sign of rectification at a cold metal probe, or of the thermoelectric force. Experiments of this kind form an accurate method

of measuring diffusion coefficients, the necessary concentrations being so small that a good approximation to infinite dilution is obtained. Some diffusion coefficients obtained in this way are shown in table 2. The diffusion coefficient D_j is expressed as

$$D_j = D_0 \exp(-\epsilon_j/kT). \qquad \ldots\ldots(5.7)$$

'Lattice defect' refers to the acceptors found in heat-treated germanium. The energy necessary to remove an atom from the lattice to an interstitial position is so high that the production of lattice vacancies in this way is very improbable. In general, vacancies can only be introduced into the interior of the semiconductor by diffusion from the surface. Such defects then behave like any other diffusing impurity, and junctions can be made simply by heating and quenching an n-type sample. Both these acceptors and those due to zinc have ionization energies much greater than the nearly constant value found for acceptors of Group III, which may point to some similarity.

Junctions prepared by diffusing p- and n-type impurities from opposite faces of a thin, relatively pure sample have the advantage that near the junction the impurity gradient can be small, giving a high Zener voltage (§ 5.3.2), whereas immediately beyond the junction it can be large, minimizing the ohmic series resistance R_1. Thus such a junction can be made to have a low forward resistance and yet withstand high inverse voltages, properties which are mutually exclusive in a linear-transition junction.

Table 2

Element	Sb	As	In	Zn	Sb	Zn	'Lattice defect'
D_0 ($m^2\,sec^{-1}$)	7×10^{-5}	7×10^{-5}	4×10^{-6}	9×10^{-6}	10^{-3}	6×10^{-5}	2×10^{-6}
ϵ_j (ev)	2·2	2·2	2·2	2·2	2·5	2·5	0·53
Reference*	F	F	F	F	D.B.	D	F.T.R.

* D = Dunlap 1952, D.B. = Dunlap and Brown 1952, F = Fuller 1952, F.T.R. = Fuller, Theuerer and van Roosbroeck 1952.

Finally, it is probable that junctions can also be made by segregation. Nearly all impurities are more soluble in liquid germanium than in the solid phase but boron (acceptor) is an exception. On progressively cooling a melt containing boron and a donor impurity such as arsenic, a transition from p to n can occur along the cooling axis. The junction formed should be similar to those made by withdrawal techniques.

5.3. *Properties of p–n Junctions*

Two factors make p–n junctions of particular interest by comparison with other semiconductor devices. Firstly, a good measure of quantitative agreement has been found between theory and the experimental electrical properties of germanium junctions. Secondly, direct measurements of the potential distribution across the barrier region have been made. In measurements of this type, Pearson, Read and Shockley (1952) have found reasonable agreement between the observed distribution and that predicted by Shockley (1949).

The d.c. characteristic can be divided into four main parts (fig. 18). At relatively low voltages, the inverse current is nearly constant, whilst the forward current increases exponentially with voltage. These two regions correspond to the rectifier characteristic predicted from the theory of an ideal junction. At higher forward voltages, the current tends to be limited by R_1, resulting in a nearly linear variation of current with voltage. For high inverse voltages, the

current suddenly begins to increase very rapidly. This has been ascribed (McAfee *et al.* 1951) to field induced excitation of carrier pairs by the mechanism postulated by Zener (1934).

5.3.1. *Forward Characteristic*

It is difficult to compare R_1 with its theoretical value, since this requires a detailed knowledge of the impurity distribution. The exponential portion (together with the low inverse voltage region where the current I_s flows) can be fitted to Wagner's ideal rectification equation (eqn. (5.1)), provided that the effects of series resistance are minimized by measuring V between potential probes on either side of the barrier. The value of η so obtained is often within a few per cent of unity (Goucher *et al.* 1951). I_s can be obtained from the low voltage ($|V|<kT/e$) resistance, and also directly from the inverse characteristic. Both values agree well with that obtained from independent measurements of L_p, L_n, σ_p and σ_n on the same sample.

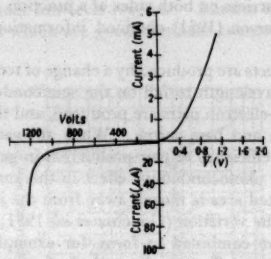


Fig. 18. Current–voltage characteristic of a p–n junction (note change of scale) (*after* Pietenpol).

As predicted, the flow of forward current is accompanied by minority carrier injection. No direct measurements of γ for junctions have been reported, but its value may be deduced in a particular case as follows. The current gain α reported for a particular n–p–n transistor (§5.4.3) is 0·89 (Shockley *et al.* 1951). Since it is thought that no current multiplication mechanism is present, γ must be at least as large as this. According to Shockley, the quantity (electron current into p-region/hole current into n-region) should be equal to $(\sigma_n/\sigma_p)^{1/2}$, giving $\gamma=[1+(\sigma_p/\sigma_n)^{1/2}]^{-1}$. Values quoted as typical for such a transistor are

$$\sigma_n=10^4\Omega^{-1}m^{-1}, \qquad \sigma_p=10^2\Omega^{-1}m^{-1},$$

so that the theoretical value for γ is 0·91.

The forward characteristic also determines the floating potential which is developed, in the same way as in point contacts, when the material forming a junction contains a non-equilibrium density of minority carriers. Detailed measurements have been made (Shockley *et al.* 1951) which are in agreement with the theory.

5.3.2. *Inverse Characteristic*

At voltages rather less than the Zener voltage, the inverse current is approximately independent of voltage. The reasons for its deviation from exact constancy

have not been discussed, but it is unlikely that any sort of ohmic leakage path is the cause, since the incremental conductance often increases nearly proportionally to the current when the latter is increased by the presence of minority carriers.

At low inverse voltages, a junction forms an excellent collector, since its transparency to minority carriers is unity, while that for normal carriers is nearly zero. It differs from the point contact in that it sets up no appreciable field, so that carriers flow to it by diffusion only. A junction transmits carriers from either side with equal facility, so that if, for example, holes in n-type material are to be detected, the conductivity of the p-type half of the junction must not be too low, lest the wanted hole current be swamped by an unwanted electron current. Current amplification like that found in point contacts is never observed with junctions.

Because of this transparency to minority carriers, a junction is sensitive to anything which changes the minority carrier densities. An increase in hydrostatic pressure widens the intrinsic gap in germanium, thus decreasing the densities of minority carriers on both sides of a junction. By utilizing this effect Hall, Bardeen and Pearson (1951) obtained information on the variation of ϵ_i with pressure.

More important effects are produced by a change of temperature or by allowing light of appropriate wavelength to fall on the semiconductor near the junction. In the latter case, hole–electron pairs are produced, and the minority components diffuse to the junction and pass across. When allowance is made for recombination, the quantum efficiency of pair production in germanium is found to be unity, as with the bulk photoconductive effect in this material. The sensitivity falls off if the illuminated area is moved away from the junction, and L_p and L_n can be deduced from the variation (Goucher *et al.* 1951).

If two junctions are combined to form, for example, an n–p–n structure, illumination has the same effect as the arrival of minority carriers at a hook collector, and quantum efficiencies much greater than unity can be obtained. It has been suggested that this structure is a model for some types of secondary photoconductivity (McKay 1951 b). Other types of ionizing radiation produce similar effects (McKay 1951 a).

The saturation current of a p–n junction increases more rapidly with temperature than does the inverse current of a point contact. The temperature dependence of I_s can be described in terms of an activation energy, and the value so obtained for germanium junctions is 0·72 ev, in good agreement with the theoretical value, which should be the intrinsic gap width, $\epsilon_i = 0{\cdot}76$ ev. No measurements of the noise component of the saturation current have been published, but the noise figures of n–p–n transistors imply that it is much less than for contact collectors.

Above a certain critical inverse voltage (the Zener voltage) the current begins to increase very rapidly. The form of the variation of this current is in good agreement with that deduced from Zener's (1934) theory of dielectric breakdown (McAfee *et al.* 1951), and the Zener voltage is, as predicted, nearly independent of temperature. The magnitude of the critical field can be calculated from the Zener voltage and the transition rate, which can be estimated from the junction capacity. It is found to be constant, independent of barrier thickness, at about 2×10^7 v m^{-1}, compared with the theoretical value of 7×10^7 v m^{-1}.

5.3.3. *Signal Properties*

As would be expected, the a.c. admittance of an inverse-biased junction is chiefly that of the barrier capacity C. If the transition is of the linear type considered by Shockley, a plot of $\log C$ against $\log V$ should give a straight line of slope $-\frac{1}{3}$, whereas for an abrupt transition from p- to n-type conductivity the slope should be $-\frac{1}{2}$. Measurements on junctions made by withdrawal give a slope between $-\frac{1}{2}$ and $-\frac{1}{3}$, indicating that the actual transition is intermediate between the two types (Pearson *et al.* 1952).

The admittance near zero bias depends on frequency in a way which shows that it cannot be represented by a parallel combination of conductance and capacity. This admittance varies in the way which would be expected if it were due to minority carrier currents, and from a plot of its real and imaginary parts against frequency (fig. 19), a value of minority carrier lifetime can be obtained which is in agreement with that found optically. It will be seen that for frequencies such that $\omega\tau > 2$, the conductance and acceptance are nearly equal, so that the junction is a poor low-level rectifier. Similar effects occur also at high voltages, and junctions which are good rectifiers cannot be used at frequencies above about 50 kc/s.

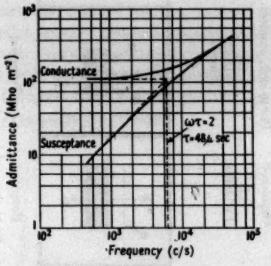


Fig. 19. Admittance of a p–n junction as a function of frequency (*after* Goucher *et al.* 1951).

5.4. *Technological Applications*

The electrode arrangements of some representative semiconductor circuit elements are shown in fig. 20. Most of them depend on minority carrier flow and they are conveniently classified according to the number of elements, type of carrier flow, and type of active element (contact or junction).

5.4.1. *Diodes*

Diodes are two-electrode elements which provide an asymmetrical current–voltage characteristic for circuit applications. Their properties are, of course, just those of the contact or junction they comprise.

Contact Diodes. Germanium contact diodes make use of the rectifying contact between a sharply pointed metal wire attached to one external lead, and the surface of a germanium pellet which is attached to the other lead by a large area soldered contact. The wire contact, which is pointed by grinding or by an

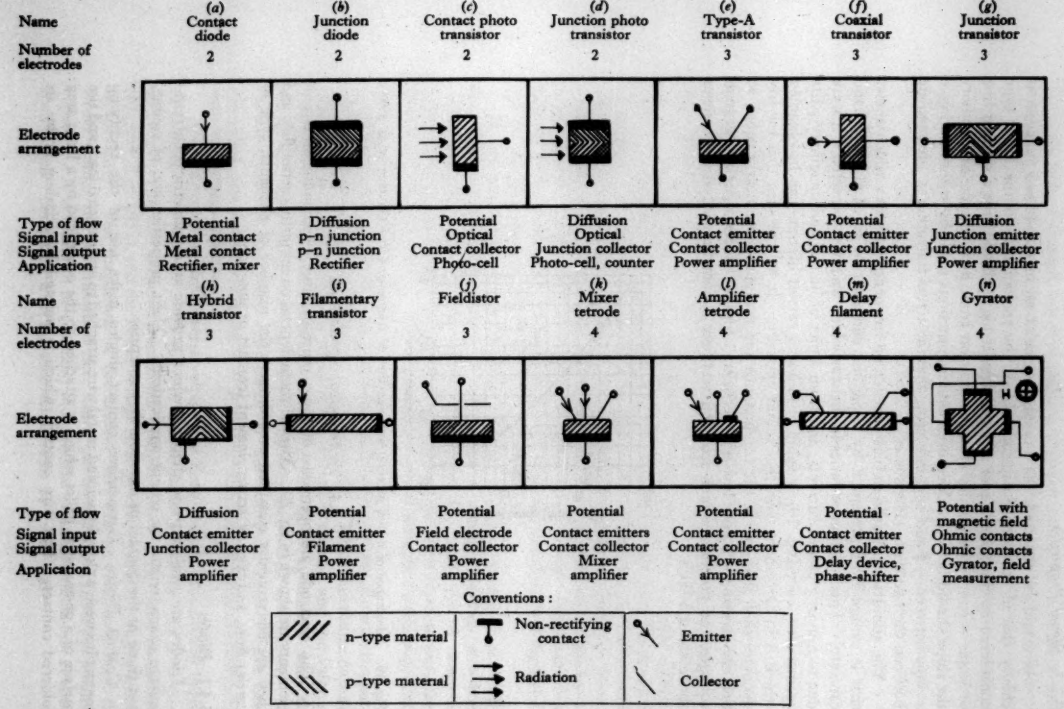

Name
(a) Contact diode
(b) Junction diode
(c) Contact photo transistor
(d) Junction photo transistor
(e) Type-A transistor
(f) Coaxial transistor
(g) Junction transistor
Number of electrodes
2
2
2
2
3
3
3
Electrode arrangement
Type of flow
Signal input
Signal output
Application
Potential
Metal contact
Metal contact
Rectifier, mixer
Diffusion
p–n junction
p–n junction
Rectifier
Potential
Optical
Contact collector
Photo-cell
Diffusion
Optical
Junction collector
Photo-cell, counter
Potential
Contact emitter
Contact collector
Power amplifier
Potential
Contact emitter
Contact collector
Power amplifier
Diffusion
Junction emitter
Junction collector
Power amplifier
Name
(h) Hybrid transistor
(i) Filamentary transistor
(j) Fieldistor
(k) Mixer tetrode
(l) Amplifier tetrode
(m) Delay filament
(n) Gyrator
Number of electrodes
3
3
3
4
4
4
4
Electrode arrangement
H
Type of flow
Signal input
Signal output
Application
Diffusion
Contact emitter
Junction collector
Power amplifier
Potential
Contact emitter
Filament
Power amplifier
Potential
Field electrode
Contact collector
Power amplifier
Potential
Contact emitters
Contact collector
Mixer amplifier
Potential
Contact emitter
Contact collector
Power amplifier
Potential
Contact emitter
Contact collector
Delay device, phase-shifter
Potential with magnetic field
Ohmic contacts
Ohmic contacts
Gyrator, field measurement
Conventions :
n-type material
p-type material
Non-rectifying contact
Radiation
Emitter
Collector

Fig. 20.

electrolytic process, is pressed against the etched surface of the pellet by a spring, usually a part of the wire itself. This assembly is mounted in an insulating capsule, which is then hermetically sealed. The capsule may also be partly or wholly filled with an inert compound to provide mechanical stability and further resistance to the harmful ingress of moisture.

The properties of some typical commercial germanium diodes are shown in table 3. It will be seen that some compromise between the various quantities has to be made according to the purpose for which the diode is intended.

Table 3.

Type of diode	General purpose	Low imped. h.f. mixer	High inv. voltage	High impedance
Turnover voltage (v)	100	20	200	100
Forward current at +1v (mA)	5	20	4	2
Inverse current at −10v (μA)	20	100	10	5
Maximum frequency (Mc/s)	50	1000	50	50

The capacity of the complete unit is usually 0·5–1 pF. The maximum working temperature for most applications is about 70° C, at which the inverse current may have increased by a factor between 2 and 4, and the forward current by 10 to 20%.

Germanium contact diodes are now produced commercially in large quantities. They find application in nearly every branch of electronic engineering, for example as linear rectifiers up to 100 Mc/s, square-law detectors and mixers up to 1000 Mc/s, as d.c. restorers, as gates, switches etc., in pulse circuits, and as modulators and demodulators in telecommunication equipment.

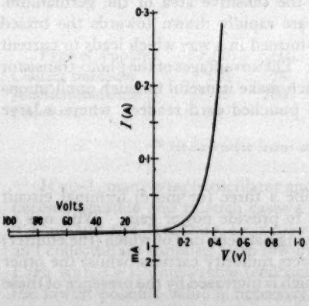


Fig. 21. Current–voltage characteristic of a p–n junction power rectifier (note change of scale) (*after* Davies, 1952).

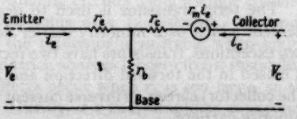


Fig. 22. Equivalent circuit of a transistor.

Junction Diodes. In junction diodes a small piece of semiconductor containing a rectifying junction is enclosed in a sealed metal can or a plastic bead. Non-rectifying connections are made to the p- and n-type portions. These diodes are not commercially available at the time of writing, so that little can be said about applications and ratings, but they are expected to find wide application as power supply rectifiers for electronic apparatus, for which the ability to withstand inverse

voltages up to 1000v and to pass mean currents up to 0·5A is required. A characteristic of a prototype germanium power rectifier is shown in fig. 21, from which it will be seen that it has the outstanding advantage over all other rectifiers of developing only a very low voltage, and hence having a very small power dissipation, during the flow of forward current densities as large as $10^7 Am^{-2}$. As the inverse current can also be made rather less than for most contact diodes, junction diodes may also have small-signal applications.

5.4.2. *Photocells and Counters*

Because of their great sensitivity to light and ionizing radiations, both point contacts and p–n junctions using germanium are useful as photocells and particle counters. The arrangements shown in figs. 20(*c*) and 20(*d*) differ from the corresponding simple diodes only in that radiation is allowed to fall on the region near the barrier. The sensitivity to radiation from the visible range up to $\lambda \simeq 1{\cdot}7$ microns is much greater than would be possible by utilization of the bulk photoconductive effect in germanium. For wavelengths greater than 1·7 microns, germanium is insensitive to radiation, and the group of compounds PbS, PbSe, and PbTe provide one of the most sensitive means for the detection of radiation in the near infra-red, using the secondary photoconductive effect of thin microcrystalline films (Smith 1951).

One of the most interesting photo-elements is the 'photo-transistor' (Shive 1949b). In one form of this, a non-rectifying 'base' contact is made to the periphery of a thin germanium disc, and a rectifying contact made to the centre of one face, whilst light is allowed to fall on the other face. The disc can be mounted inside a capsule, one end of which contains a small moulded plastic lens which serves to focus incident light on the sensitive area of the germanium. Minority carriers liberated by radiation are rapidly drawn towards the biased collector contact, which is usually electro-formed in a way which leads to current multiplication effects ($\alpha' > 1$) at the contact. The advantages of the photo-transistor are its small size and high sensitivity, which make it useful for such applications as binary data transmission systems and punched card readers, where a large number of cells is required.

5.4.3. *Transistors*

The term transistor is used to describe a three (or more) terminal circuit element which uses minority carrier flow to provide power gain. With one or two exceptions, transistors have two rectifying barriers, one of which (the emitter) is biased in the forward direction and injects minority carriers, whilst the other (the collector) carries an inverse current which is increased by the presence of these minority carriers.

Contact Transistors. The earliest and simplest form of the transistor was the so-called 'type-A', shown diagrammatically in fig. 20(*e*). In their original experiments, Bardeen and Brattain (1948, 1949) used n-type germanium, and this is still the most widely used material, although the effect has been demonstrated in p-type germanium (Pfann and Scaff 1950), silicon, and lead sulphide.

In the type-A arrangement, a non-rectifying contact is made to the base of a small block of the semiconductor, whilst two rectifying contacts are made to the opposite face of the block. This face is usually etched in order to remove an undesirable worked layer from the surface, and the rectifying contacts are made

with two sharply pointed wires. One of these, which acts as the emitter, is biased forwards by a few tenths of a volt during operation, and may be formed in order to raise the minority carrier injection ratio γ to as large a value as possible. The collector, which is placed about 50 microns away, has its resistance reduced by heavy forming in the inverse direction during manufacture. Thus the collector, with its normal operating bias of about -20 v, sets up an intense field of potential flow which ensures that nearly every hole injected at the emitter reaches it. The forming also makes possible a high value of the quantity α', as discussed in §5.1.2.

A number of alternative geometrical arrangements have been suggested for the point contact transistor, all employing the same physical principles as those of the type-A arrangement. One example is that of the coaxial transistor (fig. 20 (*f*)).

The electrical properties of a transistor may be described in terms of an equivalent circuit such as that shown in fig. 22. The incremental quantities r_e, r_c, and r_b can be regarded as passive resistances in series with the three leads, and the transistor action as resulting from an internal generator with an output $r_m i_e$. The current gain α, which is nearly equal to the current multiplication factor α' of the collector, is given by

$$\alpha=(r_m+r_b)/(r_c+r_b). \qquad \ldots\ldots(5.8)$$

In practical cases r_e and r_b are both much less than either r_c or r_m, and this fact makes a large power gain possible. The properties of circuits containing transistors can be derived from the equivalent circuits, and several detailed treatments have been given (Wallace and Pietenpol 1951, Ryder and Kircher 1949). Only a few of these properties affect the present discussion.

Table 4

Type of amplifier	$r_e(\Omega)$	$r_b(\Omega)$	$r_c(\Omega)$	$r_m(\Omega)$	Noise (db)*	f_α (Mc/s)*
Contact transistor	120	75	$1{\cdot}5\times10^4$	$3{\cdot}8\times10^4$	45	7
Junction transistor	25	250	$5{\cdot}0\times10^6$	$4{\cdot}7\times10^6$	15	5
Fieldistor	10^7	—	2×10^8	10^7	60	0·1

* Noise = noise factor at 1 kc/s; f_α = cut-off frequency.

If $\alpha>1$, many simple oscillator and switching circuits using a single transistor become possible and in such cases a large value of α is desirable. Secondly, if $\alpha>1$ and $r_b(r_b+r_m)>(r_e+r_b)(r_c+r_b)$, the transistor is unstable when emitter and collector are supplied from low-impedance sources (short-circuit instability). Even with $\alpha\not>1$, undesirable feedback effects occur when r_b is too large, so that the lowest possible value is necessary. Thirdly, if $\alpha\simeq1$, amplifier connections in which base, emitter or collector is the common electrode between input and output circuits become analogous to the grounded-grid, grounded-cathode, and cathode-follower connections respectively of a thermionic triode. Finally, in order to obtain large power gains it is, of course, necessary to make r_c and r_m large, and r_e and r_b small.

Equivalent circuit parameters of a typical germanium contact transistor are given in table 4. Some control over these is necessary to produce a useful amplifier, but details of the methods employed commercially have not yet been published. r_e and r_b can be controlled to some extent by choice of material

conductivity, whilst r_b can be greatly reduced by increasing the electrode spacing, although this impairs the frequency response (Slade 1950). α and r_e are controlled chiefly by forming processes but some compromise is necessary in the extent of these. If the collector is only lightly formed r_e is large, so that the withdrawal field is small at normal voltages; α then tends to be small both for this reason and because no 'hook' structure (or trapping region) is produced. With optimum forming r_e is reduced, but the field increases and the 'hook' appears so that α is increased. Still heavier forming causes both α and r_e to decrease. In addition to the inevitable spreading due to diffusion the non-isochronous flow field between emitter and collector results in a range of hole transit times. This causes a sharp pulse to be broadened and delayed, and the gain for high-frequency sinusoidal signals to be reduced and become complex. These effects are reflected in the equivalent circuit by changes in the magnitude and phase of r_m, the changes in the other quantities, due to the reactances of leads and barriers, being small by comparison. The magnitude of the effects is conveniently described by the value of f_0, the frequency at which r_m has fallen to $1/\sqrt{2}$ of its low-frequency value. The observed phase shift agrees with that calculated from the estimated transit time (Bardeen and Brattain 1949).

The flow can be made more nearly isochronous by applying a strong transverse magnetic field which forces the lines of current flow nearer the surface. A field of about 5×10^5 amp turns m^{-1} may raise f_0 by a factor of about 8 (Brown 1949). Similar effects can be obtained in an arrangement, sometimes called the transistor tetrode (fig. 20(*l*)), where an auxiliary current passes through a non-rectifying contact near the collector, increasing the magnitude of the flow field and making it more uniform (Reich *et al.* 1951). These effects show, incidentally, that the primary cause of the reduction in r_m at high frequencies is field spreading, and not diffusion. Further support to this view is given by the fact that f_0 is often proportional to the collector current, whereas if diffusion were important, a square-root dependence should be observed. Increasing the electrode spacing would be expected to cause the field to spread further, and does in fact reduce f_0.

It appears that an adequate response up to frequencies of about 50 Mc/s can now be obtained by simple methods (Morton 1952). Still higher input frequencies can be accepted, although not amplified directly, by a contact transistor provided with two emitters (fig. 20(*k*)). The signal is applied to one while the other receives a heterodyning voltage, the mixing effect which occurs enabling the difference frequency to be amplified by the normal transistor action. The input frequencies can be much higher than the upper frequency limit of a semiconductor amplifier, provided their difference falls within the normal range (Haegele 1950).

An undesirable property of contact transistors is their large, and largely unexplained, excess noise. The noise factor varies nearly inversely with frequency, the actual rate being about 11 db per decade of frequency change.

Junction Transistors. In a junction transistor, two regions of the same conductivity type are separated by a thin 'base' layer of the opposite conductivity type (fig. 20(*g*)). Carriers injected from one of the junctions so formed will flow across the base layer by diffusion, since the current density is low, and increase the current drawn by the second junction (the collector). Non-rectifying connections require to be made to all three parts of the transistor, and for mechanical rigidity it is customary to embed these in a protective plastic bead which surrounds the entire unit.

Only germanium junction transistors are as yet available, though it is expected that similar devices using silicon will eventually be made for applications at higher temperatures. Junction transistors can be made either by withdrawal or by diffusion. In the former method the collector section is apparently grown first from n-type germanium of conductivity about 10 Ω^{-1} m^{-1}, enough p-type impurity then being added to make the conductivity of the base about 100 Ω^{-1} m^{-1}. When the base has been grown to the required thickness, sufficient n-type impurity is added to overcome the p-type impurity already present and to give the emitter section a conductivity of about 10^4 Ω^{-1} m^{-1}. The n–p–n structure, rather than p–n–p, is probably chosen because it eases the difficult mechanical problem of making a non-rectifying contact to the base layer, which may be only about 30 microns thick. Little is known about the manufacture of transistors by diffusion, which usually have the p–n–p arrangement, although they are known to be under active development.

The properties of the three sections must be closely controlled whichever method of manufacture is used. The base and collector sections must be of fairly low conductivity material so that the saturation current of the collector junction is small, whilst the emitter section should have a high conductivity in order to suppress unwanted carrier injection into it from the base. The transition rate of the collector junction must be low enough to enable it to withstand the applied inverse voltage, and the base layer must have a certain optimum thickness to obtain the best frequency response (Shockley *et al.* 1951).

The manufacture of germanium ingots containing suitably disposed impurities is not easy, and attempts have been made to circumvent this problem by using a junction for the collector and a point contact for the emitter. This is the so-called 'hybrid' transistor (fig. 20(*h*)).

Junction transistors also can be represented by the equivalent circuit of fig. 22 (Wallace and Pietenpol 1951). In n–p–n transistors, α can be made very nearly unity, but never greater, so that short-circuit instability never occurs. The departure of α from unity is probably a result of the impossibility of making the emitter junction γ exactly unity, although volume recombination in the base layer may be a contributory factor. α increases very slightly with increasing emitter current, and it has been suggested that this indicates that the electron lifetime in the base increases with electron density. When α is nearly unity further small increases provide large returns in circuit flexibility. It will be seen from table 4 that r_e and r_b have about the same values as for modern contact transistors, while r_c and r_m are much greater, so that larger power gains are inherently possible. As would be expected, r_e is inversely proportional to emitter current, while r_c and r_m vary in the same sense but more slowly. The emitter and collector capacities are larger than for point contacts, and may limit the frequency response in some types of circuit. A more fundamental limit to the frequency response is set by the decrease of α with increasing frequency, due to the diffusive nature of the carrier flow. It is readily shown that the cut-off frequency is given by

$$f_0 = D_n/\pi W^2 \qquad \ldots\ldots(5.9)$$

where W is the thickness of base layer. Improvement can be obtained by making the base thinner, but there comes a point at which its transverse resistance is comparable with the impedances of the junction capacities. An appreciable part of the applied voltage then appears as a resistive drop in the base, and α is again

reduced because less conduction current flows into the emitter junction. There is thus an optimum thickness, apparently about 30 microns for units of useful size, which can be reduced only by decreasing the junction area.

Noise in junction transistors has a similar spectrum to that of contact types, but a much smaller amplitude.

5.4.4. *Other Circuit Elements*

The 'Fieldistor'. The 'fieldistor' amplifier (Stuetzer 1950) makes use of an unexplained variation in the current to a collector contact when an intense electric field is applied normal to the germanium surface in its neighbourhood. This field can be set up by an insulated electrode (fig. 20(*j*)), so that the input impedance and current gain can be high. The power gain and frequency response are poor by comparison with other semiconductor amplifiers (table 4).

Filamentary Transistors and Delay Lines. In the filamentary transistor (Shockley, Pearson and Haynes 1949) a contact or junction emitter injects carriers into a thin filament, along which they are drawn by an applied field, modulating its resistance (fig. 20(*i*)). The power gain is relatively low, and, because of the long transit time, the frequency response is poor. The current gain can approach $b+1$ for n-type material, and the noise level is lower than that of early type-A transistors. It may have applications where a low output impedance is required.

It might be thought that some use could be made of the signal delay provided by such a structure by taking the output, not from the filament resistance, but from a contact collector placed at the end opposite to the emitter (fig. 20(*m*)). A pulse of carriers might then be injected at one end, flow along the filament, and be detected some time later by the collector at the other end. For practical pulse delay applications it is necessary to make M, defined as $M=$(delay time)/(half-amplitude width of output pulse), as large as possible. It is not difficult to make the flow field nearly isochronous, so that diffusion is the limiting factor. Reduction of the effects of diffusion requires high electric fields and consequently, since the permissible power dissipation is limited, the use of material with a low conductivity σ. If the other quantities remain fixed, $M \propto \sigma^{-1/2}$. Unfortunately the low values of σ required bring into prominence the effect, discussed in § 3.1.2, of the carrier concentration on the local conductivity. This results in a further broadening of the output pulse, which can be reduced only by decreasing the signal level. A limit is set to this by the high noise level of a contact collector, and only a compromise value of σ can be found. It can be shown by an approximate analysis that, with input pulses of 1·5 μsec duration, $M \simeq 6$ for germanium delay lines, with normal cooling, a result which has been confirmed experimentally. In applications where the low resolution represented by this value is admissible, the delay line forms a useful circuit element. It is probable that improvements can be made by reducing the noise, i.e. by using a junction collector, or by increasing the power dissipation, using forced cooling.

Magnetometers and Gyrators. It is an obvious extension of Hall coefficient measurements to use a specimen with a known coefficient to measure magnetic fields (Pearson 1948). Germanium is a particularly suitable substance, because the high mobility enables large coefficients to be combined with relatively small resistivities. A more interesting application is found when we note that a structure such as that shown in fig. 20(*n*), in which an output voltage is derived from the input current by means of the Hall effect, has some of the properties of an ideal

gyrator (Hogan 1952). An ideal gyrator is a four-terminal network in which V_1, I_1, the input voltage and current are connected with V_2, I_2, those at the output, by the relations

$$V_1 = -SI_2, \quad V_2 = SI_1. \qquad \ldots\ldots(5.10)$$

It is a passive, linear network for which the reciprocity theorem does not hold. Amongst other curious properties it has the effect of changing the phase of a signal by π for one direction of transmission but not for the other. By combining it with a normal network a filter can be constructed which transmits a signal in one direction only.

REFERENCES

AIGRAIN, P., 1950, *C.R. Acad. Sci., Paris*, **230**, 62.

BARDEEN, J., 1947, *Phys. Rev.*, **71**, 717 ; 1950, *Bell. Syst. Tech. J.*, **29**, 469.

BARDEEN, J., and BRATTAIN, W. H., 1948, *Phys. Rev.*, **74**, 230 ; 1949, *Ibid.*, **75**, 1208.

BARDEEN, J., and PFANN, W. G., 1950, *Phys. Rev.*, **77**, 401.

BARDEEN, J., and SHOCKLEY, W., 1950, *Phys. Rev.*, **80**, 72.

BENNETT, A. I., and HUNTER, L. P., 1951, *Phys. Rev.*, **81**, 152.

BENZER, S., 1944, N.D.R.C., Report 14-342, Purdue University; 1949, *J. Appl. Phys.*, **20**, 804.

BETHE, H. A., 1942, M.I.T. Radiation Lab. Report, 43/12.

BILLIG, E., and RIDOUT, M. S., 1951, *Nature, Lond.*, **167**, 1028.

BLAKEMORE, J. S., 1952, *Elect. Comm.*, **29**, 131.

BLOCH, F., 1928, *Z. Phys.*, **52**, 555.

BRATTAIN, W. H., and BARDEEN, J., 1948, *Phys. Rev.*, **74**, 231.

BRATTAIN, W. H., and PEARSON, G. L., 1950, *Phys. Rev.*, **80**, 846.

BRAY, R., 1946, Purdue University Progress Report, Contract No. W-36-039, SC 32020, June 1–Aug. 31, 1946; 1949, *Phys. Rev.*, **76**, 152.

BRAY, R., LARK-HOROVITZ, K., and SMITH, R. N., 1947, *Phys. Rev.*, **72**, 530.

BROWN, C. B., 1949, *Phys. Rev.*, **76**, 1736.

CHASMAR, R. P., and PUTLEY, E. H., 1951, *Semi-Conducting Materials*, ed. H. K. Henisch (London : Butterworths Scientific Publications), p. 208.

CONWELL, E., and WEISSKOPF, V. F., 1950, *Phys. Rev.*, **77**, 388.

DAVIES, L. J., 1952, *B.T.H. Activities*, **23**, 77.

DAVYDOV, B. I., 1938, *J. Tech. Phys., U.S.S.R.*, **5**, 87.

DILWORTH, C. C., 1948, *Proc. Phys. Soc.*, **60**, 315.

DOUGLAS, R. W., and JAMES, E. G., 1951, *Proc. Inst. Elect. Engrs.*, Pt. III, **98**, 157.

DRUDE, P., 1900, *Ann. Phys., Lpz.*, **1**, 566.

DUNLAP, W. C., 1952, *Phys. Rev.*, **85**, 945.

DUNLAP, W. C., and BROWN, D. E., 1952, *Phys. Rev.*, **86**, 417.

DUNLAP, W. C., and HENNELLY, E. F., 1948, *Phys. Rev.*, **74**, 976.

ERGINSOY, C., 1950 a, *Phys. Rev.*, **79**, 1013; 1950 b, *Ibid.*, **80**, 1104.

ESTERMANN, I., and FONER, A., 1950, *Phys. Rev.*, **79**, 365.

FAN, H. Y., 1950, *Phys. Rev.*, **78**, 808 ; 1951, *Ibid.*, **82**, 900.

FAN, H. Y., and BECKER, M., 1951, *Semi-Conducting Materials*, ed. H. K. Henisch (London : Butterworths Scientific Publications), p. 132.

FULLER, C. S., 1952, *Phys. Rev.*, **86**, 136.

FULLER, C. S., THEUERER, H. C., and VAN ROOSBROECK, W., 1952, *Phys. Rev.*, **85**, 678.

GISOLF, J. H., 1949, *Physica*, **15**, 825.

GOUCHER, F. S., 1950, *Phys. Rev.*, **78**, 646.

GOUCHER, F. S., PEARSON, G. L., SPARKS, M., TEAL, G. K., and SHOCKLEY, W., 1951, *Phys. Rev.*, **81**, 637.

GUBANOV, A. I., 1950, *J. Tech. Phys., U.S.S.R.*, **20**, 1287.

GUNN, J. B., 1952, *Proc. Phys. Soc.* B, **65**, 908; 1953, *Ibid.*, B, **66**, 330.

HAEGELE, R. W., 1950, *Onde Elect.*, **30**, 239.

HALL, H. H., BARDEEN, J., and PEARSON, G. L., 1951, *Phys. Rev.*, **84**, 129.

HALL, R. N., and DUNLAP, W. C., 1950, *Phys. Rev.*, **80**, 467.

HAYNES, J. R., and SHOCKLEY, W., 1949, *Phys. Rev.*, **75**, 691 ; 1951, *Ibid.*, **81**, 835.

HAYNES, J. R., and WESTPHAL, W. C., 1952, *Phys. Rev.*, **85**, 680.
HENISCH, H. K., 1949, *Metal Rectifiers* (Oxford : University Press).
HERRING, C., 1949, *Bell. Syst. Tech. J.*, **28**, 401.
HOGAN, C. L., 1952, *Bell. Syst. Tech. J.*, **31**, 1.
HOGARTH, C. A., and GRANVILLE, J. W., 1951, *Proc. Phys. Soc.* B, **64**, 992.
HUNG, C. S., 1950, *Phys. Rev.*, **79**, 727.
HUTNER, R. A., RITTNER, E. S., and DU PRÉ, F. K., 1950, *Philips Res. Rep.*, **5**, 188.
JOHNSON, V. A., and LARK-HOROVITZ, K., 1947, *Phys. Rev.*, **71**, 374; 1950, *Ibid.*, **79**, 176.
JOHNSON, V. A., SMITH, R. N., and YEARIAN, H. J., 1950, *J. Appl. Phys.*, **21**, 283.
KOCK, W. E., and WALLACE, R. L., 1949, *Elect. Eng.*, **68**, 222.
LARK-HOROVITZ, K., 1951, *Semi-Conducting Materials*, ed. H. K. Henisch (London : Butterworths Scientific Publications), p. 47.
LARK-HOROVITZ, K., MIDDLETON, A. E., MILLER, E. P., and WALERSTEIN, I., 1946, *Phys. Rev.*, **69**, 258.
LEHOVEC, K., ACCARDO, C. A., and JAMGOCHIEN, E., 1951, *Phys. Rev.*, **83**, 603.
LORENTZ, H. A., 1905, *Proc. Amsterdam Acad.*, **7**, 438 ; 1909, *Theory of Electrons* (Leipzig : Teubner).
MCAFEE, K. B., RYDER, E. J., SHOCKLEY, W., and SPARKS, M., 1951, *Phys. Rev.*, **83**, 650.
MCAFEE, K. B., SHOCKLEY, W., and SPARKS, M., 1952, *Phys. Rev.*, **86**, 137.
MACFARLANE, G. G., 1947, *Proc. Phys. Soc.*, **59**, 366 ; 1950, *Ibid.* B, **63**, 807.
MCKAY, K. G., 1951 a, *Phys. Rev.*, **84**, 829 ; 1951 b, *Ibid.*, **84**, 833.
MEACHAM, L. A., and MICHAELS, S. E., 1950, *Phys. Rev.*, **78**, 175.
MEYERHOF, W. E., 1947, *Phys. Rev.*, **71**, 727.
MONTGOMERY, H. C., and SHOCKLEY, W., 1950, *Phys. Rev.* **78**, 646.
MOORE, A. R., and HERMAN, F., 1951, *Phys. Rev.*, **81**, 472.
MORTON, J. A., 1952, *Bell. Syst. Tech. J.*, **31**, 411.
MOTT, N. F., 1939, *Proc. Roy. Soc.* A, **171**, 27.
OHL, R. S., 1952, *Bell. Syst. Tech. J.*, **31**, 104.
PANTCHECHNIKOFF, J. I., 1950, *Phys. Rev.*, **79**, 1027.
PEARSON, G. L., 1948, *Rev. Sci. Instrum.*, **19**, 263.
PEARSON, G. L., and BARDEEN, J., 1949, *Phys. Rev.*, **75**, 865.
PEARSON, G. L., READ, W. T., and SHOCKLEY, W., 1952, *Phys. Rev.*, **85**, 1055.
PEARSON, G. L., STRUTHERS, J. D., and THEUERER, H. C., 1949, *Phys. Rev.*, **75**, 344 ; 1950, *Ibid.*, **77**, 809.
PFANN, W. G., 1951, *Phys. Rev.*, **81**, 882.
PFANN, W. G., and SCAFF, J. H., 1949, *Phys. Rev.*, **76**, 459; 1950, *Proc. Inst. Radio Engrs.*, **38**, 1151.
PRIM, R. C., 1951, *Bell. Syst. Tech. J.*, **30**, 1174.
PUTLEY, E. H., 1947, *Nature, Lond.*, **160**, 710 ; 1949, *Proc. Phys. Soc.* A, **62**, 284.
RAYNOR, G. V., 1952, *Rep. Prog. Phys.*, **15**, 173 (London : Physical Society).
REICH, H. J., SCHULTHEISS, P. M., SKALNIK, J. G., FLYNN, T., and GIBSON, J. E., 1951, *J. Appl. Phys.*, **22**, 682.
VAN ROOSBROECK, W., 1950, *Bell Syst. Tech. J.*, **29**, 560.
RYDER, E. J., and SHOCKLEY, W., 1949, *Phys. Rev.*, **75**, 310; 1951, *Ibid.*, **81**, 139.
RYDER, R. M., and KIRCHER, R. J., 1949, *Bell. Syst. Tech. J.*, **28**, 367.
SCHOTTKY, W., 1939, *Z. Phys.*, **113**, 367.
SCHOTTKY, W., STÖRMER, R, and WAIBEL, F, 1931, *Z. Hochfrequ.*, **37**, 162.
SEITZ, F., 1940, *Modern Theory of Solids* (New York : McGraw-Hill).
SHIVE, J. N., 1949 a, *Phys. Rev.*, **75**, 689 ; 1949 b, *Ibid.*, **76**, 575.
SHOCKLEY. W., 1949, *Bell Syst. Tech. J.*, **28**, 435 ; 1950 a, *Phys. Rev.*, **78**, 173 ; 1950 b, *Ibid.*, **78**, 294 ; 1951 a, *Electrons and Holes in Semi-Conductors* (New York : van Nostrand); 1951 b, *Bell Syst. Tech. J.*, **30**, 990.
SHOCKLEY, W., PEARSON, G. L., and HAYNES, J. R., 1949, *Bell Syst. Tech. J.*, **28**, 344.
SHOCKLEY, W., SPARKS, M., and TEAL, G. K., 1951, *Phys. Rev.*, **83**, 151.
SITTNER, W. R., 1952, *Proc. Inst. Radio Engrs.*, **40**, 448.
SLADE, B. N., 1950, *R.C.A. Rev.*, **11**, 517.
SMITH, A. H., 1949, *Phys. Rev.*, **75**, 953.
SMITH, R. A., 1951, *Semi-Conducting Materials*, ed. H. K. Henisch (London : Butterworths Scientific Publications), p. 198.

SOMMERFELD, A., 1928, *Z. Phys.*, **47**, 1.
STUETZER, O. M., 1950, *Proc. Inst. Radio Engrs.*, **38**, 868.
SUHL, H., and SHOCKLEY, W., 1949, *Phys. Rev.*, **75**, 1617.
TAMM, I., 1932, *Phys. Z. Sowjet.*, **1**, 733.
TEAL, G. K., SPARKS, M., and BUEHLER, E., 1951, *Phys. Rev.*, **81**, 637.
THEUERER, H. C., and SCAFF, J. H., 1951, *J. Met.*, **191**, 59.
TORREY, H. C., and WHITMER, C. A., 1948, *Crystal Rectifiers*, (New-York : McGraw-Hill).
VALDES, L. B., 1952, *Proc. Inst. Radio Engrs.*, **40**, 445.
WAGNER, C., 1931, *Phys. Z.*, **32**, 641.
WALLACE, R. L., and PIETENPOL, W. J., 1951, *Proc. Inst. Radio Engrs.*, **39**, 753.
WILSON, A. H., 1931 a, *Proc. Roy. Soc.* A, **133**, 458; 1931 b, *Ibid.*, **134**, 277; 1932, *Ibid.*, **136**, 487.
YEARIAN, H. J., 1950, *J. Appl. Phys.*, **21**, 214.
ZENER, C., 1934, *Proc. Roy. Soc.* A, **145**, 523.
VAN DER ZIEL, A., 1950, *Physica*, **16**, 359.

ELECTRICAL DISCHARGES

By F. LLEWELLYN JONES
Department of Physics, University College of Swansea

CONTENTS

PAGE

§ 1. Introduction 217

§ 2. The electrical breakdown of gases in static fields 218
2.1. Breakdown at low pressures 218
2.1.1. Theory 218
2.1.2. Secondary processes 220
2.2. Breakdown at higher pressures and in the atmosphere 222
2.2.1. Secondary processes 224
2.2.2. Photo-ionization as a secondary mechanism 225
2.2.3. The formative time lag of breakdown 228
2.3. Breakdown under extreme conditions of pressure or gap distance 229
2.3.1. Highly compressed gases 229
2.3.2. Very short gaps 230

§ 3. High-frequency discharges 231
3.1. Breakdown at high frequencies 231
3.1.1. Nature of the discharge 231
3.1.2. H discharges 232
3.1.3. E discharges 233
3.2. Discharge mechanism 235
3.2.1. Breakdown 235
3.2.2. The maintained uniform glow 237
3.3. Electron energy distribution 239
3.4. Applications 241
3.4.1. Atomic collisional phenomena 241
3.4.2. Sources (of ions and radiations) 243
3.4.3. Ionospheric non-linear properties 243

§ 4. Cold electron emission in discharges 244
4.1. Occurrence 244
4.2. Investigation of the mechanism of emission 246
4.3. Role of thin surface films 249

§ 5. The regime of space charge—the discharge established 252
5.1. D.C. arc discharges 252
5.1.1. Theory of the arc column 253
5.1.2. Cathode effects 254
5.1.3. Heavy current arcs 256
5.1.4. Short arcs at electrical contacts 257
5.2. Plasma oscillations 258
5.2.1. Oscillations in glow discharges 258
5.2.2. Noise 259
5.3. Spectroscopic phenomena and ion concentrations 259
5.4. Effects of irradiation of the discharge 260

References 260

ABSTRACT. Recent developments in electrical discharges are surveyed under the main headings: The Breakdown of Gases in Static Fields; High-Frequency Discharges; Cold Emission Phenomena; and The Regime of Space Charges. Investigations of the fundamental process of electrical breakdown at high and low pressures for long and short gaps by measurements of the growth of ionization current just prior to breakdown are described; a generalized view is made possible regarding breakdown as a consequence of the development of pre-breakdown ionization currents by primary and secondary processes which are themselves controlled by the nature and geometry of the cathode. Properties of E and H high-frequency discharges are outlined, and a general theory of high-frequency breakdown and of maintained discharges is given. It is shown how further knowledge of the behaviour of ionized gases is obtained which has led to a greater understanding of collisional processes such as attachment and recombination. Various applications are indicated. Studies of the important rôle played by surface films in the cold extraction of electrons from electrodes are described, and applications of the results to arc discharges, high-voltage high-pressure breakdown, and cathode phenomena in breakdown are indicated. Work on the nature of radio-frequency plasma oscillations, and of the generation of noise by discharges, is described. An account of recent work on arc discharges is given, dealing particularly with cathode phenomena and the thermal and optical properties of the arc column.

§1. INTRODUCTION

In elucidating the varied and complex phenomena of electrical discharges it is necessary to identify the significant processes of ion generation and loss which occur, and also to investigate the energy and probability relations of those processes and the energy distribution functions of the colliding particles.

The subject of discharge physics was reviewed comprehensively in *Reports on Progress in Physics* by Francis and Jenkins (1940) and in *Reviews of Modern Physics* by Druyvesteyn and Penning (1940); certain further aspects, such as the cathode fall and spark breakdown at high pressures, were later treated in *Reports on Progress in Physics* by Lunt, von Engel and Meek (1941) and in *Advances in Physics* by Massey (1952). Fundamental phenomena such as attachment and recombination, both dissociative and associative, secondary emission, and collisions involving photon–atom interaction have been reviewed in *Reports on Progress in Physics* by Massey (1949), in *Advances in Physics* by Massey (1952), and treated comprehensively by Massey and Burhop (1952). Reference should be made to all these previous publications.

In recent years considerable developments have been made in the understanding of electrical discharges, and these reflect the increased interest in various aspects such as electrode surface phenomena, collisional and ionization processes involved in the static and high-frequency breakdown of solid, liquid and gaseous dielectrics, the precise nature of atom–ion reactions which are of particular interest in relation to the ionosphere, ion sources for use in nuclear physics, and, finally, the thermal and optical properties of arc and glow discharges. The modern trend of using increasingly high voltages in the transmission of electrical power on the large scale, as well as the use of high-voltage nuclear particle accelerators in the laboratory, has raised further problems of fundamental physical interest.

This Report surveys recent developments in electrical discharges under the following main headings:

The electrical breakdown of gases in static fields;
High-frequency discharges;
Cold emission phenomena in discharges; and
The regime of space charges—the discharge established.

This order approximately represents a range of discharge phenomena covering the basic processes of the production of electrons and the resulting growth of currents leading to breakdown and the establishment of high-current space-charge conditioned discharges. Further, the phenomena attributed to the cold emission of electrons are so widespread, and appear in so many different aspects of electrical discharges, that it is considered necessary to devote a whole section to that subject.

The physical problem involved is to explain in terms of fundamental collisional processes of electrons, atoms, ions and photons and the physics of surfaces a range of phenomena which includes, on the one hand, the micro-sparks at low-voltage electrical contacts and, on the other hand, the large sparks from high-voltage equipment. The phenomena in this wide range appear so different that different theories have, in fact, been proposed to account for the various aspects. Recent experimental and theoretical investigations show, however, that there are basic processes common to all these various aspects, thus providing the basis of a comprehensive general theory in which there is a growing recognition of the important role which can be played by electrode surface phenomena in electrical discharges.

The general problem of breakdown will be discussed here in three parts corresponding, as has previously been thought, to quite different sets of physical phenomena: breakdown at low pressures, breakdown at higher pressures, including the important case of a centimetre gap in the atmosphere, and breakdown at high voltages with highly compressed gases, thus including recent trends in high-voltage insulation.

Much recent work has also been carried out with high frequency discharges, and the resulting increased knowledge of the behaviour of ionized gases has led to a greater understanding of various collisional processes such as attachment and recombination. High-frequency discharges have found further applications as mono-chromatic ion sources for particle accelerators and light sources in spectroscopy. The investigation of radio-frequency plasma oscillation discharges is a modern development of considerable interest, and further advances have also been made in our knowledge of thermal, optical and electrical properties of arc discharges and of related cathode phenomena.

The writer wishes here to acknowledge the help he has received in preparing this *Report* from Drs. J. Dutton, S. C. Haydon, C. G. Morgan and G. C. Williams, of the Department of Physics, University College of Swansea, who have been closely associated with some of the work described below. His thanks are also due to the Royal Society for kind permission to reproduce figs. 5, 6, and 7, and to the Editors of *Nature* for fig. 4.

§2. THE ELECTRICAL BREAKDOWN OF GASES IN STATIC FIELDS

2.1. *Breakdown at Low Pressures*

2.1.1. *Theory.*

A satisfactory explanation of breakdown at low pressures is that of the generalized Townsend theory, based on measurements of the growth of a small photoelectric current between two parallel electrodes at distance d apart in a gas under an electric field X. When the field is constant, the growth of an initial current I_0 photoelectrically emitted from the cathode is given for small values of the electrode separators d by

$$I = I_0 \exp C_1 d, \qquad \ldots\ldots(1)$$

and at larger values of d by the relation

$$I = I_0 \exp C_1 d / \{1 - C_2 (\exp C_1 d - 1)\}, \qquad \ldots\ldots(2)$$

where C_1 and C_2 are constants when the field X and gas pressure p are constant. This is best illustrated by plotting $\ln I/I_0$ against d as shown in fig. 1. It is clear that by comparing the equation with the observed values of I the constants C_1 and C_2 can be found for any given value of the electric field. It should be noted that $I = 0$ when I_0 is 0; i.e. an initial current I_0 is essential for maintaining the ionization under these conditions.

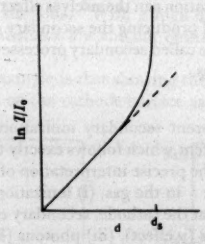


Fig. 1. Growth of ionization currents with electrode separation at low gas pressures.

The form of this relation is interesting: since in many gases C_2 is small, the $(\ln I/I_0, d)$ graph is linear for low values of d (eqn. (1)), then curves upwards as d increases until a value of d_s is reached such that the denominator vanishes. This upcurving is of vital significance: I approaches infinity when

$$1 - C_2 (\exp C_1 d - 1) = 0. \qquad \ldots\ldots(3)$$

The physical interpretation of this relation is that I remains finite even if I_0 is zero; i.e. the current continues to pass even when there is no initial ionization. The ionization current becomes self-maintained. The gas no longer insulates and the current can increase as much as the external circuit permits. Equation (3) is then the spark-breakdown criterion and d_s the spark distance. This relation in fact relates the spark parameter $p \times d_s$ and spark voltage, i.e. it gives Paschen's law. The theory can be tested by obtaining C_1 and C_2 from measurements of the growth of the current, then calculating d_s and comparing it with the value actually observed. The complete agreement found when the parameter $p \times d$ is less than about 150 cm mm Hg confirms the basic ideas of this theory.

Consequently it is generally accepted that the breakdown mechanism at these lower pressures is controlled by the coefficients C_1 and C_2 which can be measured from the growth of an ionization current under conditions remote from electrical breakdown. To understand the breakdown, therefore, it is necessary to understand the physical significance of the coefficients C_1 and C_2. That is, it is necessary to find which collision processes involving electrons, atoms, photons, ions, etc. these coefficients represent. When this is known, the coefficients themselves can then be calculated from the various atomic cross sections and energy distribution functions. The interpretation of C_1 is clear—it is Townsend's coefficient α of primary ionization by electrons, and can be represented by the expression $p f(X/p)$.

When, in addition, a magnetic field is present, this modifies the motion of the electrons and changes the conditions under which breakdown occurs. Breakdown in a tranverse magnetic field has been treated theoretically by Valle (1950, 1952) and by Somerville (1952), who obtains sparking potential curves in qualitative agreement with measured values obtained by Penning (1937). Experimental determinations of α in a magnetic field are necessary as a basis for a more thorough theoretical treatment.

However, the motion of the electrons crossing the gap and building up an avalanche produces other effects, such as excitation of the neutral molecules, and possibly their dissociation, in addition to the production of positive ions. These products of the primary ionization can themselves effect ionization, as, for example, by falling on the cathode and producing the secondary emission of electrons. For this reason these processes are called secondary processes, because they take place as a result of the primary process.

2.1.2. *Secondary processes.*

A number of quite different secondary ionization processes can produce a growth of the ionization current which follows exactly the same form as that given by eqn. (2), and this makes the precise interpretation of the coefficient C_2 difficult. These possible processes are : in the gas, (i) ionization of gas atoms in collisions with positive ions (β effect); at the cathode, secondary emission of electrons due to incidence of (ii) positive ions (γ effect), (iii) photons (δ effect), (iv) excited atoms (ϵ effect). This differs from the δ effect in that the energy required to liberate electrons from the cathode is transferred to the cathode as excited atoms instead of as photons.

Another possible secondary process is that of the photo-ionization of the gas. Experimental evidence on this subject is scanty, and in any case the values of the absorption coefficients are such that at low pressures this process is not significant compared with the other processes. It can be easily shown that approximately $C_2 = \omega/\alpha$, where

$$\omega = \beta + \alpha\gamma + \delta + \epsilon. \qquad \ldots\ldots(4)$$

Although the shape of the ionization growth curve cannot itself decide which of these secondary processes predominate in any given case, other characteristics of the processes can help to decide. For example, the form of the dependence of the coefficient ω/α on the nature of the cathode surface, and also on the mean energies of the photons and ions, can separate processes (ii) and (iii). A peak or a fall in the (ω/α, X/p) curve as X/p increases indicates that the predominant process is that of photoelectric emission from the cathode ; on the other hand, a small steady increase of ω/α with X/p indicates the predominating influence of emission due to impact of positive ions (Llewellyn Jones 1939).

As for direct evidence of these processes, little or no experimental evidence for a β effect has been found by passing positive ions through a rarified gas ; and at the moderate electric intensities found at the lower gas pressures this process is less likely than those acting at the cathode. The experiments of Penning (1928) and Oliphant and Moon (1930) have demonstrated the efficiency of γ and ϵ processes; again the photoelectric effect at electrodes due to ionization currents is well known (Townsend 1923). Further, the experimental evidence of the significance of photo-ionization in breakdown is neither extensive nor conclusive. Thus at these lower pressures the experimental evidence favours the predominance

of processes of cathode emission. Such emission, unlike gas ionization, depends on the effective work function of the cathode surface as well as on the energies of the photons or ions, and it is on this aspect that considerable work has been done in recent years (Llewellyn Jones and Davies 1951 a, b).

The general experimental procedure was to alter the effective work function of a cathode of pure metal by condensing on its surface different layers of atoms so as to produce electrical double layers, and so raise or lower the work function at will. In this way it was found that in hydrogen, for example, with a low work function cathode, the secondary emission was due mainly to the impact of low-energy photons and of positive ions, the photoelectric emission being about ten times greater than that due to the ions. With a high work function cathode surface, on the other hand, the emission is low and due entirely to the ions. Typical results are shown in fig. 2.

Another significant example is that showing the influence of thin tarnish films or layers of adsorbed gas on the cathode surface as shown in fig. 3. This shows

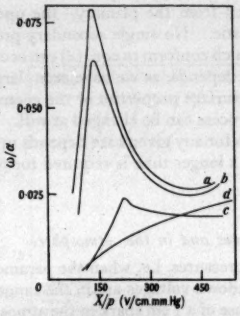


Fig. 2. The effect on ω/α of deposition of a copper film on nickel: curve *a*, thin copper film; curve *b*, film outgassed; curves *c* and *d*, increase of film thickness and outgassing.

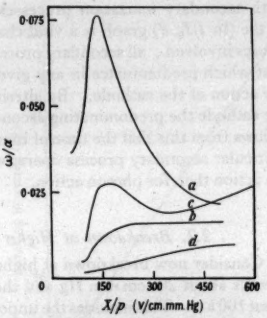


Fig. 3. Effect of surface cleaning on secondary emission (ω/α) of copper: curve *a*, slightly oxidized; curve *b*, after baking; curve *c*, after 10 hours exposure to hydrogen; curve *d*, cleaned by ion bombardment and outgassed.

how the surface films controlled the nature of the secondary ionization by making emission due to low-energy photons the predominating mechanism; when the film was removed only the high-energy positive ions produced any significant emission.

Again, writing the breakdown criterion (3) as

$$1-(\omega/\alpha)[\exp\{pdf(V_s/pd_s)\}-1]=0, \qquad \ldots\ldots(5)$$

where V_s is the sparking or breakdown potential, it can be seen that changes in ω/α produce corresponding changes in the sparking potential V_s. It was found that V_s was more than doubled by changing the cathode from pure outgassed aluminium to a cathode consisting of a layer of atoms of Ni on a pure Al base, this combination having a high work function. Results such as these show the

predominating influence of the cathode in low-pressure discharges in uniform fields.

When the field is not uniform Townsend showed originally how eqn. (2) must be modified to take into account the variation of the factors C_1 and C_2 across the gap. When C_2 represents a cathode mechanism in a discharge tube of, for example, two concentric cylinders of radii r_1 and r_2, eqn. (2) is replaced by

$$1-(\omega/\alpha)\{\exp\int_{r_1}^{r_2}\alpha' dr-1\}=0, \qquad \ldots\ldots(6)$$

where α' may be put equal to α if the spatial variation of the field is not too high. Very recent experiments (Llewellyn Jones and Williams 1953b) show that the above results for uniform fields have their counterpart in non-uniform fields, and so confirm the view that the predominating secondary processes of the breakdown mechanism are determined by the nature and geometry of the cathode.

The conclusion arrived at, therefore, is that breakdown is the result of a continuous process of ionization growth determined by primary ionization associated with secondary ionization processes derived from the primary—the upcurving of the $(\ln I/I_0, d)$ graph is a vital characteristic. No single secondary process is always involved ; all secondary processes which conform to eqn. (2) can occur, and that which predominates in any given case depends, as we have seen, largely on the action of the cathode. By altering the surface properties or the geometry of the cathode the predominating secondary process can be changed at will. It also follows from this that the time of breakdown for any given case depends upon the particular secondary process operative, e.g. a longer time is required for positive ion action than for photon action.

2.2. *Breakdown at Higher Pressures and in the Atmosphere*

Consider now breakdown at higher gas pressures, i.e. when the parameter *pd* exceeds about 200 cm mm Hg and the breakdown voltages are in the range 10 to about 100 kv. This includes the important case of a 1 cm spark in the atmosphere.

Up to about 1925 or so the above theory was thought to apply also to these high pressures, mainly due to lack of evidence to the contrary, but the experiments of Rogowski showed that impulse sparks in the atmosphere could take place in times of the order of 10^{-7} second. Positive ions could not cross the gap in this time. This result therefore threw doubt on the theory of breakdown based on positive ions crossing the gap, and this in turn gave rise to experiments to test whether there was in fact any secondary ionization at all occurring under these conditions (Paavola 1929, Masch 1932, Sanders 1932, Hochberg and Sandberg 1942).

The procedure in all these experiments was the same : the growth of an ionization current under a uniform field was measured for different values of the distance d between the plates. The results, plotted as $(\ln I/I_0, d)$ curves, showed that at these high pressures the graphs were entirely linear and broke off suddenly when a spark occurred : i.e. there was apparently no curved portion similar to those found at low pressures. Two important conclusions were drawn from this (Loeb and Meek 1941, Loeb 1939) : (i) The straight lines indicated the presence of the primary ionization process (C_1) (this process alone could not produce a spark). (ii) The appearance of a spark before any curvature in the graph was found indicated the

sudden appearance of an entirely new mechanism of breakdown not dependent on a secondary ionization process (C_2) as outlined above : i.e. there was no C_2.

Thus breakdown in these conditions appeared to present an outstanding problem.

For these reasons the spark criterion based on the Townsend idea of primary and secondary ionization was widely rejected in favour of an entirely different theory of breakdown based on positive ion space charges. These space charges were considered to distort the field and assist photo-ionization at the head of an electron avalanche, the whole process occurring extremely rapidly (Loeb and Meek 1941, Meek 1940, 1942, Loeb and Wijsman 1948). This view was unfortunately never put into quantitative form to relate the breakdown potential to the ionization processes of photo-ionization and electron collision. Also, no indication was given in this new view as to the conditions when the earlier Townsend theory ceases to apply and when the space charge theory takes control, but it was stated that when pd_s exceeds 150 cm mm Hg, and certainly for a 1 cm gap in the atmosphere ($pd = 760$ cm mm Hg) the new theory applied. Further, since breakdown was con-

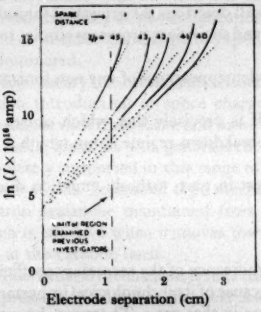


Fig. 4. Growth of photoelectric currents in air at 200 mm of mercury pressure.
(*Reproduced by courtesy of 'Nature'*)

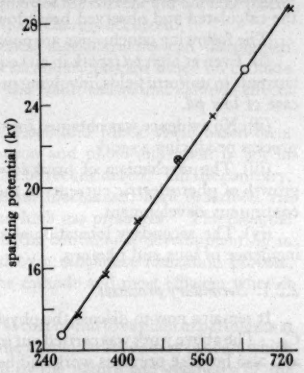


Fig. 5. Calculated (x) and observed (o) sparking potentials in air.
(*Reproduced by courtesy of the Royal Society*)

sidered to be a gas process only, the cathode had no influence on the mechanism. No experimental evidence has been given to indicate whether these processes, such as the suggested influence of photo-ionization and space charges, do in fact act in the way proposed and, if they do, whether the orders of magnitude are right. Theoretical considerations suggested the contrary (Zeleny 1942). In particular, no quantitative analysis of the influence of photo-ionization on the growth of the electron avalanche was made : the suggested influence of this action was based on descriptive and qualitative considerations only (Bruce 1951). Since, however, the basis of this new view was derived from the previous experimental measurements of ionization currents at high pressures (e.g. Sanders 1932), it is essential to analyse very carefully the conclusions drawn from this work.

When this is done it can be shown that sufficient attention had not been paid either to the requisite scope of those measurements themselves or to their accuracy. The upcurving can only occur when d is large enough for the factor $(\omega/\alpha)(\exp \alpha d - 1)$ to be experimentally measurable; the previous work did not extend to sufficiently large gap distances because the anode potential was not maintained steady enough to allow the current at those large distances to be measured.

For reasons such as these the growth of currents in the crucial region just prior to the appearance of a spark has recently been investigated (Llewellyn Jones and Parker 1950, 1952). The results given in fig. 4 showed that over a restricted distance the graphs were linear, as previous workers had found; when, however, d was increased to the crucial region near breakdown the graphs were found to curve up, contrary to previous work, and were exactly as found for low pressure.

To test the spark criterion (3) from these growth of current curves, C_1 was found from the straight parts using eqn. (1), and C_2 from the curved portions using eqn. (2), and these were substituted in eqn. (3) to find d_s theoretically. After each set of runs a single but extremely careful determination of V_s was made for comparison. These values are given in fig. 5, which shows full agreement between the calculated and observed breakdown potentials.

The following conclusions are now drawn:

(i) Even at high pd (spark in air) eqn. (2) still describes the growth of ionization currents in uniform fields, involving primary and secondary processes similar to the case of low pd.

(ii) No evidence was obtained for the sudden appearance of any new ionization process producing a spark.

(iii) The mechanism of breakdown itself is precisely that which causes the growth of photoelectric currents in the pre-breakdown regime, from which it is a continuous development.

(iv) The secondary ionization was, at least in part, cathode emission due to incidence of ions and photons.

2.2.1. *Secondary processes.*

It remains now to discuss the physical significance of the secondary coefficient C_2. The above work was carried out in air because of its technological importance, and also because previous work had been done in that gas. There are, however, two important reasons for carrying out similar experiments in a simple gas like nitrogen. First, air can produce changes in the cathode surface due to formation of oxides under ion impact; secondly, air being a mixture of gases of different ionization potentials, photo-ionization might take place because excitation of nitrogen molecules can produce photons capable of ionizing oxygen. Clearly, in a simple non-active gas these two effects cannot occur. Experiments (Dutton, Haydon and Llewellyn Jones 1952 a, b) were therefore carried out in nitrogen in order to investigate

(1) whether an upcurving of the $(\log I, d)$ graph occurred in nitrogen as well as in air—to see whether breakdown in air (a mixture) is fundamentally different from that in a simple gas because of the possibility of photo-ionization;

(2) whether the breakdown potential or the secondary coefficients depended upon the current, i.e. whether double ionization processes or photon production by recombination processes occurred; and

(3) whether space charges had any significant action in breakdown.

The use of nitrogen involved further experimental difficulties because of the necessity of increasing the sensitivity of the current measurements by a hundred times, and the voltages between 10 and 40 kv had to be steady to within 0·002% : an electrostatic balance capable of measuring currents of 3×10^{-13} A with anode potentials of 10 kv had to be developed. Finally a careful investigation of the variations of photoelectric properties over the cathode surface was required.

Results are given in fig. 6 which give growth of current curves for different values of X/p but the same initial current I_0. The curves show that nitrogen, like air, also gives an upcurving following precisely eqn. (2) and, contrary to previous views, this disposes of the suggestion that photo-ionization is a process essential to breakdown under these conditions. Figure 7 gives the growth of current curves up to breakdown for three different initial currents. These show that the same sparking potential is obtained for widely different electron currents (a total range of 10^6 : 1) ; and further, that over the wide current range examined the same sparking potential was obtained. This showed that the secondary ionization did not depend on the current, and that therefore the production of photons by recombination processes depending on the square of the current did not take place. Further, no significant effects attributable to space charges occurred.

The results indicate that the experimental measurements can distinguish between the effect of photo-ionization and a secondary process based on cathode effects : in these parallel plate experiments the result shows that cathode effects predominated.

The theory of breakdown (Loeb 1939, Loeb and Meek 1941) based on a sudden introduction of space charge distortion and photo-ionization is not in accordance with these facts, and must therefore be reconsidered. On the contrary, the basic ideas of the generalized Townsend mechanism here described are completely supported in this range of pd_s and high gas pressures. The essential feature of this Townsend type mechanism is the continuous development of an electron avalanche maintained from behind by a secondary ionization process, which is effective when it moves towards the cathode and most efficient when it acts at the cathode itself.

This general mechanism is thus the same as that found to explain experiments at low pressures, and it would appear also that the particular secondary process which predominates in any given case is influenced by the nature and geometry of the cathode.

After the state represented by the criterion (3) is reached, i.e. break down, the current can rise at a rate determined by the external circuit, and generally this involves high current densities, intense ionization, and the transition to an arc, and a space charge conditioned discharge follows.

2.2.2. *Photo-ionization as a secondary mechanism.*

Since it has been concluded from experiment that photo-ionization played no significant part in breakdown in a uniform field in nitrogen, it would appear reasonable to conclude that the results also applied to air, although in that gas a mechanism for the production of ionizing photons does exist. However, detailed mathematical investigation of the growth of ionization currents in air on the basis of photo-ionization treated as a secondary process has very recently been made by Dutton, Haydon and Llewellyn Jones, and Davidson (1953a). This analysis, which can only be summarized here, shows that photo-ionization as a secondary ionization

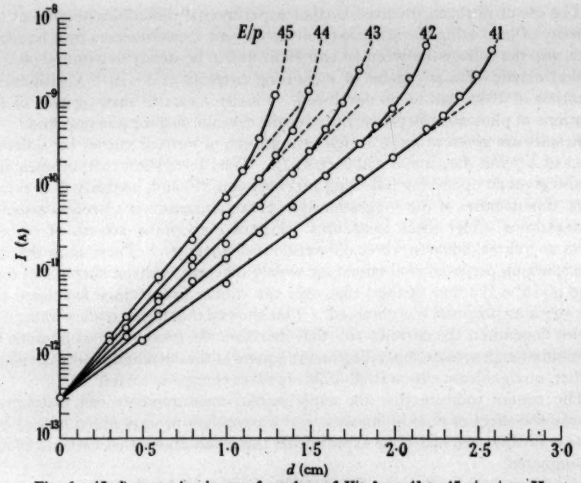


Fig. 6. (I, d) curves in nitrogen for values of X/p from 41 to 45 v/cm/mm Hg at a pressure of 300 mm.
(*Reproduced by courtesy of the Royal Society*)

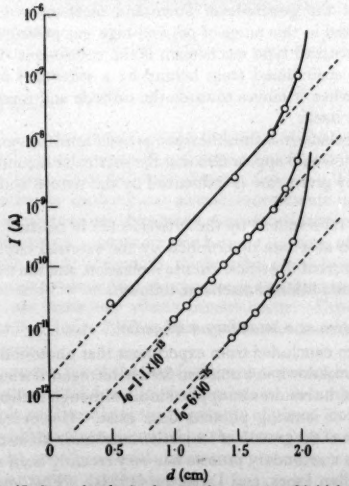


Fig. 7. (I, d) graphs obtained over a large range of initial currents I_0.
(*Reproduced by courtesy of the Royal Society*)

process can lead to breakdown, but that this can occur only under certain conditions which impose restrictions on the values of the various photon absorption coefficients. In this respect, therefore, photo-ionization is different from the other secondary processes involving the β, γ, or δ effects. In these latter cases, no matter what value is assigned to the secondary coefficents, it is always possible to obtain from the sparking criterion (3) a distance at which breakdown will occur. This is not the case with photo-ionization. Thus evidence that photo-ionization can occur in a gas is not in itself evidence that this process can always produce breakdown. The mathematical analysis also shows that when, as in any practical case, resonance fluorescence and degeneracy occur, then the absorption coefficient of the photons which lead to ionization of the gas atoms must be small ($\sim 3\ \text{cm}^{-1}$ at 200 mm Hg). Simple calculations then show that a large number of high-energy photons will reach the cathode so that in, say, a 1 cm gap in air it is to be expected that the cathode itself will have the dominating influence on the mechanisms leading to breakdown. There is evidence from recent experimental investigations into this particular aspect of breakdown that this is in fact the case. One other conclusion follows from these recent calculations. On account of the very complicated relation between the various probability factors of excitation, absorption and ionization, which obtain in the growth of the ionization current, qualitative considerations of the action of photons in any particular case of breakdown can be not only inaccurate but also misleading. It is advisable that the growth of the ionization current be calculated for the particular gas and electrode configuration under consideration before concluding whether photo-ionization is, or is not, the significant secondary mechanism. In order to make such calculations, even for the simplest case, there is at present considerable need for accurate experimental determinations of the various atomic cross sections for photo absorption and for excitation by electron impact.

These considerations are relevant to the interesting case of breakdown in highly non-uniform fields, when the point or smaller electrode is the anode and the cathode is at a great distance. These conditions, in the way indicated in § 2.1, practically eliminate those secondary processes which occur at the cathode, and thus leave only the gas processes of ionization by positive ions or by photon absorption. Now earlier investigations showed that the ionizing power of comparatively slow positive ions was extremely small and required onset energies of the order of 10^2 ev, and for this reason photo-ionization was considered to be the only possible gas ionization process. Recent experiments, however, have thrown new light on this difficult question. Horton and Millest (1946) devised a sensitive method of detecting the ionization of neutral atoms of helium by fast neutral atoms, and it was shown that the earlier techniques, especially those depending on space charge effects, were far too insensitive to detect the likely ionization produced by positive ions. Horton and Millest found that neutral atoms of helium ionized other neutral atoms of helium when their energy exceeded 50 ev ($= 2\,V_1$) exactly as predicted by simple collision theory. This result is very different from the figure of 300 ev previously given for the minimum energy of the colliding atom for ionization. This result supports the possibility of positive ions being able to ionize (Werner 1950, Medicus 1948) by a process, for example, in which the ions are first neutralized, then ionize as a fast atom.

In the present state of knowledge these two secondary processes acting in the gas appear, on general grounds, to be possible secondary ionization processes

capable of producing breakdown for this particular configuration of a positive point discharge. However, as pointed out above, qualitative discussion of the relative importance of ionization by photons or by positive ions is of little significance : the question can only be decided by accurate measurement of the pre-breakdown ionization currents and detailed quantitative calculation, as has been done for uniform fields. This has still to be done for non-uniform fields.

2.2.3. *The formative time lag of breakdown.*

The agreement between the observed static sparking potential and that calculated by inserting the coefficients α and ω/α (obtained by experiments on the growth of ionization currents in the steady state when $V<V_s$) in eqn. (3) shows that the breakdown of a gas is a consequence of the action of primary and secondary ionization processes of the Townsend type. However, if the applied voltage V is greater than the static sparking potential V_s (or if $V=V_s$ and there is also a maintained I_0) a non-steady state exists and the ionization current then increases with time. Under these conditions the current grows until the current density becomes sufficient to give a glow discharge and a rapid fall of the potential across the gap. In practice, the time elapsing between the appearance of an initiatory electron in the gap and the commencement of the potential fall is usually defined as the formative time lag. This formative time lag depends on the various ionization processes involved and on the overvoltage, which is defined as the amount by which V exceeds V_s. At low values of pd Schade (1937) obtained agreement between his observed formative time lags and those calculated by first obtaining an expression for the temporal growth of ionization current in terms of the primary and secondary ionization coefficients and then making the assumption that breakdown was complete when the current from the cathode reached a certain fixed value. The only secondary ionization mechanism which Schade found it necessary to consider was that of positive ion action at the cathode.

Detailed theoretical calculations of the dependence of the formative time lag on overvoltage, taking into account the two secondary ionization processes of secondary electron emission from the cathode due to the incidence of positive ions and photons, have recently been made (see Davidson's appendix, Dutton, Haydon and Llewellyn Jones 1953 b) for a 1 cm gap in air at atmospheric pressure. The details cannot be given briefly here, but it is of interest to consider the method adopted and some of the results obtained. A rigorous expression for the temporal growth of the ionization current in terms of the coefficient of secondary ionization ($\omega/\alpha=\gamma+\delta/\alpha$) was first obtained by integrating the continuity equation following the general procedure used by Bartholomcyczyck (1940) and von Gugelburg (1937). Schade's condition that breakdown is complete when a certain fixed current leaves the cathode was assumed for simplicity. Since such particular values for this current have not been measured, calculations of the formative time lag were made using this current as a parameter. In this way a wide range of cases of practical interest was covered.

The results of the analysis show that the same primary and secondary ionization processes which give rise to the observed growth of pre-breakdown currents given by eqn. (2) under static conditions with a uniform field lead to a rapid decrease of the formative time lag with increasing overvoltage. In

fact, time lags as low as 10^{-6} sec can be obtained with an overvoltage as low as 2% when the significant secondary process is photoelectric emission from the cathode. These short formative lags at low overvoltages are a consequence of the extremely rapid development of ionization due to the ordinary primary and secondary ionization processes when the applied potential exceeds the static sparking potential, a fact not sufficiently appreciated previously. Thus the experimentally observed short formative time lags are consistent with a theory of breakdown dependent on primary and secondary processes of the Townsend type, and these short time lags in overvolted gaps do not therefore necessitate some other breakdown mechanisms to account for them as has previously been supposed (Loeb and Meek 1941). This conclusion is consistent with the conclusions of Kachickas and Fisher (1952) following their very recent measurements of the formative time lags in nitrogen (see also Ganger 1949, and Loeb 1951).

The theoretical investigation of Dutton, Haydon and Llewellyn Jones, and Davidson (1953 b) also showed how it is possible by careful analysis of experimental determinations of the variation of time lag with overvoltage to obtain information concerning the relative importance of the various possible secondary ionization processes. In particular, comparison of the results of this analysis with the experimental measurements of Fisher and Bederson (1951) indicates that the predominant secondary process under their particular experimental conditions was probably photoelectric action at the cathode together with a small amount of emission by positive ions, a result which is consistent with the results of experiments on the growth of the pre-breakdown ionization corrents (Llewellyn Jones and Parker 1950).

2.3. *Breakdown under Extreme Conditions of Pressure or Gap Distance*

2.3.1. *Highly compressed gases; high pd.*

Of the remaining groups of breakdown phenomena, that at very high voltages and very high gas pressures is of particular interest owing to the modern use of high gas pressures for high-voltage insulation. This application is based on the well-established Paschen law which states that the sparking potential is a function only of the parameter pd. Therefore to insulate a high voltage V a conveniently short distance d may be employed provided the gas pressure is correspondingly high. In this way high-voltage equipment may be made conveniently small. Recent work (Trump, Cloud, Mann and Hanson 1950), however, with comparatively short gaps (of the order of centimetres) and very high pressures ($\sim 10^2$ atmospheres), showed that the breakdown potential was lower than that expected from Paschen's law. It is now necessary to consider the breakdown mechanism in these circumstances.

Paschen's law is based on the fact that the fundamental collisional processes must be functions of the ratio X/p. If, however, any of the processes of ionization were to depend on the factors X or p separately, then the Paschen relationship would not hold. A clue to an explanation of the failure of Paschen's law at very high pressures can be found by comparing, for example, values of the electric field ($X = V/d$) for the case of breakdown in the atmosphere in a 1 cm gap with the case of a gap of only 0·1 mm in air at a pressure of a 100 atmospheres, the parameter X/p being the same in both cases. In the 1 cm gap the field is 3×10^4 v/cm, while

in the shorter gap at 100 atmospheres the field is $\sim 10^6$ v/cm, and with such high fields electrons can be extracted from the cathode by a field process, thus increasing the ionization in the gaps and so lowering the breakdown potential. The theory of breakdown can thus be generalized to include this field emission. Because the new ionization depends on X and not on the ratio X/p, Paschen's law does not hold. Cold emission under such conditions is discussed in more detail in § 4. It has recently been found that high cold cathode emission can occur in these circumstances of moderately high fields of the order of 10^5 v/cm, and this is consistent with the observed failure of Paschen's law (Llewellyn Jones and Morgan 1951).

For these reasons the full practical advantages expected from the adoption of high-pressure gas insulation are not always obtained. The factor of importance is the nature of the surface of the cathode when the field is high—this is especially the case for modern gas-insulated cables.

Another aspect of the breakdown of gases at high values of the spark parameter pd is the case of a very long gap at atmospheric pressure. This differs from the case just considered of very high p and small d in that the average electric field is much lower (Meek and Perry 1946, Bruce 1948). Since a very long spark in the atmosphere is a particular case of breakdown in a non-uniform field, it should be noted that the discussion in § 2.2.2 is relevant. Considerable experimental data obtained under controlled conditions, as well as detailed information of the meteorological conditions, are required, however, before a precise physical picture of the basic electrical processes in, for example, the lightning discharge can be set up.

2.3.2. *Very short gaps.*

The last group in this survey is that which includes the micro-discharges which occur, for example, at the closing contacts of a telephone relay. Germer and Haworth (1948, 1949) have examined the production of minute discharges which occur at potentials of about 50 v even when the distance between electrodes is no greater than 10^{-4} cm. This fact at once raises a difficulty when attempting to explain the breakdown: such short distances may not permit sufficient electron amplification $\exp \alpha d$ in the gap. The explanation put forward by Germer and Haworth is that there may be a highly efficient electron extraction mechanism from the cathode, and they have shown that positive ions can play an important part by setting up a high field at the cathode (see § 5.1.4 below). An extraction mechanism involving the incidence of positive ions on surface films has independently been investigated by Llewellyn Jones, de la Perrelle and Morgan (1951), and is found to be a field dependent process (see § 4) (Llewellyn Jones and Morgan 1953, Llewellyn Jones 1952, 1953 a, b). Further investigation of this type of breakdown is still required, but it can be seen that the breakdown mechanism at low voltages and micro-distances can involve the same kind of processes which determine breakdown at the other end of the scale with very high voltages and very high gas pressures.

In concluding this brief survey of recent work on the breakdown mechanism, it would appear that in the generalized view of breakdown dependent on primary and secondary ionization processes, here applied over a very wide range of experimental facts, no single secondary ionization process is always responsible; a numberof different secondary mechanisms can and do occur, sometimes together, sometimes singly; and the factors which mainly determine the predominant

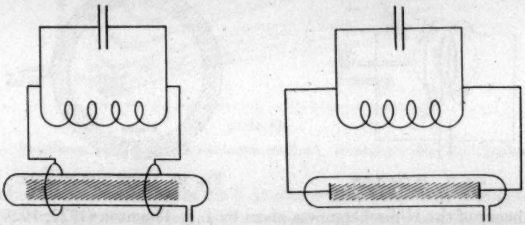
process in any particular case are the nature of the cathode surface and geometry of the system. In this respect, therefore, the physics of the electrical breakdown gases under these conditions has much in common with the physics of surfaces.

§3. HIGH-FREQUENCY DISCHARGES

3.1. *Breakdown at High Frequencies*

3.1.1. *Nature of the discharge.*

In oscillatory electric fields free electrons in a gas can acquire energies sufficient to excite and ionize the gas. The resulting motion of the electrons (and possibly ions) produces an oscillatory current which increases with the applied field until breakdown occurs and the gas becomes luminous. The oscillatory electric field may be produced in the gas of a discharge tube by connecting either internal or external electrodes to a source of high-frequency potential, or by induction from a conductor carrying an oscillatory current. Under static fields there is a continual stream of electrons or ions to one or other of the electrodes, and the electrons and ions are thus removed from the discharge. This loss is in addition to that occurring at the walls or in the gas, so that the electric field must be high enough to generate electrons and ions to make good this loss. In high-frequency fields, on the other hand, this loss to the electrodes can be reduced to insignificance, so that high-frequency discharges can occur at values of the field very much lower than those



(*a*) Electrodeless discharge. (*b*) Discharge with internal electrodes.

Fig. 8. High-frequency discharges.

required with static fields. Such properties make the high-frequency discharge of particular interest in many applications. High-frequency discharges have not been previously reviewed in these *Reports*, but various aspects have been recently surveyed by Morgan (1953) and by Prowse (1950).

Discharges are conveniently classified according to the mode of excitation (Babat 1947):

(i) The E or electrostatic discharge, which is excited by the electric field between two conductors as shown in fig. 8. Over a large range of pressures the glow discharge is similar to the positive column of a direct-current discharge in appearance and mechanism (Townsend 1928, Johnson 1930, Llewellyn Jones 1931).

(ii) The H, or ring, discharge where the discharge current forms a closed path in the gas and the glow appears as a luminous ring (Hittorf 1884) as shown in fig. 9.

If ψ is the magnetic flux through an area S of the gas, the exciting field E_r is given by

$$\oint E_r\,ds = \int_S \dot{\psi}\,dS.$$

The mechanism of the E and H discharges is fundamentally the same, and division into the two types is only justified when the wavelength λ of the high-frequency field is large compared with the linear dimensions d of the discharge; when $\lambda \simeq d$ the two forms co-exist.

3.1.2. *H discharges.*

The interpretation of observations is usually complicated by the simultaneous occurrence of both the E and the H types, the E discharge due to the alternating potential difference between the ends of the coil and the H discharge due to the induction field. In general, initiation of the discharge is mainly due to the electrostatic field, and the H discharge appears later with more intense excitation (Townsend and Donaldson 1928, MacKinnon 1929, Yarnold 1932).

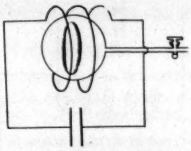


Fig. 9. H discharge.

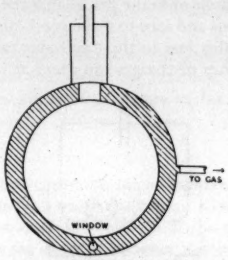


Fig. 10. Torus used by Ware.

A theory of the H discharge was given by J. J. Thomson (1927, 1928) on the assumption that an electron acquires sufficient energy in the induced electric field to ionize in a single free path. The curves relating the maximum value of the magnetic field and the gas pressure are of the same form as the (V_s, pd) curves for static field discharges, and the pressure at which the magnetic field amplitude is a minimum will be a function of the frequency. Breakdown, however, is influenced by the presence of the E field (Yarnold 1932), which must be eliminated if measurements on the H discharge are to have any quantitative value. Because of the experimental difficulties involved in the study of the H discharge alone (there are no electrodes between which the current and the potential differences may be measured, and also the length of the discharge path cannot be estimated with any degree of accuracy), the H discharge has been comparatively little studied. Certain characteristics of the discharge, for example the induced electromotive force and the current, have, however, been measured (Knipp and Knipp 1931, Tykocinski-Tykociner 1932, Smith 1941, 1947, Ware 1951).

Ware (1951) and Cousins and Ware (1951) have recently examined H discharges in helium, krypton, argon, oxygen, nitrogen and hydrogen contained in a glass torus coated on the outside with a layer of copper through which was discharged

1800 joules from a bank of condensers. The experimental arrangement is indicated in fig. 10. Estimates of the ion concentration were made from the broadening of the Balmer lines. Oscillations of the discharges appeared to be excited by large pinch effect forces found with the high currents ($\sim 10^4$ A) used.

Smith (1941, 1947) has estimated the mean electron energies and concentration using probe methods: forbidden lines were excited at high electron energies.

3.1.3. *E discharges.*

Variation of breakdown potential with frequency. At low frequencies and high gas pressures the breakdown process is basically the same as with static fields. The transit times of the electron cloud between the electrodes are short compared with the period of the field, and the breakdown potential is then approximately equal to that for steady applied potentials. For higher frequencies of continuous oscillations there is a progressive lowering of V_s (Reukema 1927, 1928, Lassen 1931), and this is explained by the incomplete removal of the electrons and ions from the gap during a half-period of the field. It is under these conditions that what is generally referred to as the high-frequency discharge occurs.

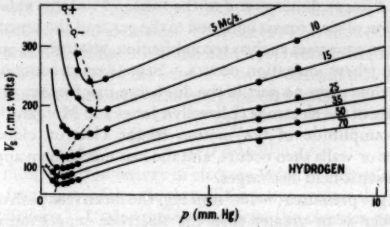


Fig. 11. Breakdown in tube A with electrodes oxidized. [Points •— after de-oxidation.]

A number of detailed studies of the E discharge have been made with different electrode systems, pressures, frequencies and gases. Gutton and Gutton (1928) examined hydrogen discharges in cylindrical tubes in the range 50 kc/s to 90 Mc/s; Hayman (1929) investigated breakdown in helium and neon for the range of frequency from 1 to 7·5 Mc/s with external sleeve electrodes, and Llewellyn Jones (1931) investigated the uniform electrodeless glow in helium at a frequency of 6·6 Mc/s: Thomson (1930, 1934, 1937) studied hydrogen for the range 2·8 to 90 Mc/s, and Rohde (1932) oxygen for the range 31 to 70 Mc/s, both using internal plane electrodes. Recent measurements of V_s by Llewellyn Jones and Morgan (1951 a, b) in air and hydrogen, and by Llewellyn Jones and Williams (1953 a) in helium in a coaxial electrode system have been made over the range of frequency from 5 to 70 Mc/s. The variation of breakdown potential with frequency for continuous oscillations for a non-uniform field is shown in fig. 11.

The ratio X/p diminishes as p increases, just as with static field discharges. When the frequency f exceeds about 10^9 c/s a state is attained, depending on the ratio f/p, where the period of the field approaches the transit time of the electrons between collisions with the gas molecules. Under these conditions the ionizing efficiency of the field is reduced and V_s increases with the frequency (Hale 1948). Experiments of Cooper (1947) and Posin (1948) on the breakdown mechanism of

pulsed microwave discharges showed that the breakdown field strength is dependent on the extent to which the gap is irradiated. Cooper measured the field at breakdown in both rectangular and coaxial waveguides at pressures from 20 to 760 mm Hg and with wavelengths from 3 to 10 cm with a pulse width of 1 μsec at a repetition rate of 400/second. The (X_s, p) curve was found to be linear over this range, unlike the (V_s, p) curve for the same system under static fields. Cooper showed that the breakdown field for the high-frequency discharge at atmospheric pressure was 70% of the value under static fields, and was independent of the width of the waveguide and the wavelength. Prowse and Jasinski (1951) have also examined the onset and rate of progress of breakdown of gases under extra high frequency conditions in cavity resonators.

Influence of boundaries. Gutton and his collaborators (1923, 1924, 1928) found under certain conditions two minima in the (V_s, p) curve, and these were shown by Gill and Donaldson (1931) to be produced when the amplitude of the oscillations of the cloud of electrons enabled them to reach the walls of the tube or the electrodes. Thus the influence of the walls or the electrodes on the discharge mechanism depends upon the relative values of the gas pressure p, oscillation frequency f, and linear dimension d of the tube. For large values of these parameters the motion of electrons is confined to the gas, and the ionization is balanced by the loss due to processes such as recombination, attachment and diffusion away from the region where ionization occurs. Secondary production of electrons at the boundaries thus plays no part in the discharge mechanism, and V_s is independent of the nature of the electrode (Llewellyn Jones and Morgan 1951 a). As p or f is reduced, the amplitude of the motion of the electron cloud increases, loss to the electrodes or walls then occurs, and the loss mechanism approximates more to that found in static field discharges.

At very low gas pressures ($\sim 10^{-3}$ mm Hg) the mean free path of the electrons is of the same order as or greater than the electrode separation or length of the discharge tube, and the chance of a collision between an electron and a gas molecule is low. Under these conditions the experiments of Gill and von Engel (1948, 1949) and of Chenot (1948) show that a discharge may be initiated at potential differences (less than 100 v) which do not depend on the gas pressure provided the frequency is greater than a certain critical value. The experimental results are interpreted in terms of the emission of electrons from the walls of the tubes by electrons which gain energy during a half-oscillation of the field. The energy which the electrons must gain from the field in order to eject electrons from the glass is of the order of 90 v. The electrons ejected from the glass wall are then accelerated by the field, and, provided the phase relationship is correct, they will arrive at the opposite wall in a half-period of the field. The ejection of the electrons from the walls then results in a rapid increase in the number of electrons which excite the gas, resulting in a visible glow. Owing to the high electron energy the nature of the gas and small variations in the pressure have very little influence on the onset potential. For each spacing of the walls a cut-off frequency is observed below which it is impossible to obtain breakdown with even very high fields ($\sim 10^2$ v/cm). Hatch and Williams (1951) have recently shown that above the cut-off frequency there is a frequency range in which two values of the breakdown field are obtained. Preliminary experiments of von Engel and Francis (1950) appear to show that a positive space charge reduces the amplitude of the electron oscillation at these low pressures.

3.2. *Discharge Mechanism*

3.2.1. *Breakdown.*

The mechanism of the high-frequency discharge is, in general, simpler than that of the static field discharge since there is no uni-directional drift of electrons away from the body of the gas, thereby necessitating a secondary source of electrons to replace the primary electrons lost to the electrodes. The mechanism of the discharge in its simplest form then consists of two processes: the generation of electrons, usually by single electron impact with gas molecules (α process) and loss of electrons by diffusion, and also, under certain conditions, by drift, attachment or recombination. The necessary conditions are that the mean free path l should be small compared with a linear dimension a of the tube, which is itself small compared with the wavelength λ, and the oscillation frequency f must be less than the collisional frequency ν of electrons with gas molecules. When electron attachment can be neglected and recombination is small, breakdown occurs when the significant processes are those of electron generation by ionization by collision and electron loss by diffusion and drift.

When $f<\nu$ electrons acquire energy over a number of free paths, and ionization and excitation will occur (Townsend 1928, 1933, Heidemann 1931) as the mean energy of the electron cloud increases. Holstein (1946 a) showed that for a uniform field between parallel plate electrodes the field X_s at breakdown, which occurs when the rate of ionization exceeds the rate of loss by diffusion, is related to the gas pressure p and the electrode separation d by the expression

$$(pd)^2=\frac{\pi^2 E}{(\alpha/p)(X_s/p)}, \qquad \ldots\ldots(7)$$

where E is the mean electron energy in electron volts. This relation holds when f is less than the frequency of the elastic collisions but greater than the frequency of the inelastic collisions. Under these conditions the electron energy distribution is the same as that in a static field equal to the r.m.s. value of the high-frequency field (Townsend and Llewellyn Jones 1931 a, b, Townsend 1932, 1933, Holstein 1946 b), and the values of α, E and drift velocity W obtained from static field experiments may then be used. These considerations have been confirmed (Krasik, Alpert and McCoubrey 1949) in argon at pressures from 3 to 100 mm Hg when $f=3000$ Mc/s. Margenau (1946) showed, however, that in general the efficiency of the primary ionization process was not only dependent on the values of the field X and the pressure p but also on the oscillation frequency f. Herlin and Brown (1948 a, b) then showed that if q is the volume rate of production of electrons per electron, n_- the electron concentration and D_- the coefficient of diffusion, then the breakdown criterion is expressed by

$$D_-\nabla^2 n_- + qn_- = 0. \qquad \ldots\ldots(8)$$

When n_- is zero at the electrodes, the solution for uniform fields gives the sparking condition

$$q/D_-X^2=\xi=\pi^2/d^2X^2, \qquad \ldots\ldots(9)$$

where ξ is the high-frequency ionizing coefficient which is a function of X/p and f/p. Since

$$q=\alpha\omega_-X, \qquad \ldots\ldots(10)$$

where ω_- is the electron mobility, and

$$\omega_-/D_-=1/E, \qquad \ldots\ldots(11)$$

it follows that (9) is the same as (7). Values of ξ for air have been calculated from the measured values of X_s in uniform fields with continuous oscillations at about 3000 Mc/s, and the breakdown conditions for the non-uniform fields between coaxial cylinders are then predicted (Herlin and Brown 1948 a). The results are in good agreement with the measured values of the fields at breakdown, thus indicating that processes such as recombination, attachment and secondary emission of electrons from surfaces play a negligible part in the discharge mechanism in air. The limits within which the diffusion theory of breakdown applies have been fully discussed by Brown and MacDonald (1949).

Many earlier investigators (Rohde 1932, Thomson 1930, 1934, 1937, Kirchner 1925, Brasefield 1931) attempted to account for the breakdown field strength by assuming that the electron acquired ionizing energy over one free path, but the maximum energy gained by a free electron is, in general, small compared with that required for ionization. The variation of X_s with frequency deduced from that assumption was not in accordance with experiment, as indeed might have been expected. In hydrogen at 1 mm Hg the collisional frequency is about 10^9 collisions/sec, so that the assumption of freely oscillating electrons may be made only if the field frequency is greater than 10^9 c/s, i.e. greater than ten times the highest frequency used by Thomson (1930, 1934, 1937) for example.

On the other hand, when measurements are made at values of f/p such that $f \sim \nu$ the diffusion theory fails and the theory of the freely oscillating electron can be expected to hold, and Hale (1948) has shown that the measured and calculated values of V_s in argon and xenon agreed to within a few per cent. These measurements were made for frequencies from 5 to 50 Mc/s and at gas pressures from 2 to 5×10^{-2} mm Hg. The mean free path l used in the calculations was the classical kinetic theory value which is independent of the electron energy. In argon the value of l is of the same order as the kinetic theory value for energies greater than 6 ev, and only varies appreciably with electron energy at low energies (<6 ev). The value used by Hale is, therefore, of the correct order of magnitude.

Further information about the mechanism of the discharge may be obtained from a consideration of the similarity theorem. This relates the high-frequency breakdown potentials V_s of geometrically similar electrode systems by the expression (Margenau 1948 a, Llewellyn Jones and Morgan 1951 a)

$$V_s = \phi\,(pa, f/p). \qquad \ldots\ldots(12)$$

The theoretical basis of this similarity principle is that the effective collisional processes are functions of X/p (Townsend 1915, Holm 1924). Fundamental collisional processes involved in electrical discharges can be regarded as forming two groups: a group which is dependent on the parameter X/p, where X is the electric field, and a group which is not dependent on the parameter X/p. Processes of excitation or ionization by single electron impact are typical of the first group since the mean electron energies are dependent on X/p: liberation of electrons from electrodes by field emission, volume recombination of ions and electrons, and ionization by collisions of the second kind are processes typical of the second group. The validity of the similarity theorem for the case of similar discharge systems in any given gas can thus give information about the nature of the fundamental processes both of electron generation and loss which occur. In hydrogen and air (Llewellyn Jones and Morgan 1951), and in pure helium (Llewellyn Jones and Williams 1953 a), over a range of pressure from 0·1

to 20 mm Hg and frequencies from 5 to 70 Mc/s and with tube dimensions a of the order of a centimetre, the similarity theorem was obeyed. This result showed that the collisional processes of electron generation and loss were functions of X/p, and were therefore ionization by single impact and loss by diffusion and drift; at suitable values of f and p the effect of drift was shown to be negligible. All investigations of the breakdown mechanism support the view that space charge effects play a negligible part in breakdown.

When the helium contained small traces of impurity ($<0{\cdot}001$ %) the similarity theorem did not hold (Llewellyn Jones and Williams 1953 a). On the other hand, when the impurity was removed and the helium was as pure as possible the theorem was satisfied. The failure of the similarity theorem for similar discharge systems indicated immediately that the breakdown mechanism in impure helium involved a process, whether of ion generation or loss, which was not dependent on X/p. A possible ionization process satisfying the necessary conditions is the ionization of molecules of impurity in collisions of the second kind with metastable atoms of helium: the process can be operative when small traces of impurities, of hydrogen for example, are present. The reaction may be expressed by

$$\mathrm{He_{met}} + \mathrm{H_2} \rightarrow \mathrm{He} + \mathrm{H_2^+} + \mathrm{e}, \qquad \ldots\ldots(13)$$

where $\mathrm{He_{met}}$ represents the metastable atom of helium ($V_{\mathrm{met}} = 19{\cdot}77$ and/or $20{\cdot}4$ ev), $\mathrm{H_2}$ the normal hydrogen molecule ($V_{\mathrm{I}} = 15{\cdot}6$ ev), which was the most likely impurity in the experiments. Evidence for such collisions of the second kind has been obtained by Penning (1931) for neon containing less than $0{\cdot}002$% of argon; this lowered the breakdown potential by a large factor. Kruithof and his collaborators (see Massey and Burhop 1952) considered that the chance of the occurrence of the resonance reaction

$$\mathrm{Ne_{met}} + \mathrm{A} \rightarrow \mathrm{Ne} + \mathrm{A^+} + \mathrm{e} \qquad \ldots\ldots(14)$$

was about unity per collision. The energy transfer in (13) is large, and its efficiency correspondingly lower; if, however, the hydrogen molecule is dissociated then the excess energy may be absorbed and the process made more probable.

These results show that the similarity theorem may be used as a tool to investigate the relative importance of various collisional phenomena in discharges. In this way it should be possible not only to investigate the importance of processes such as (13), but also those involving, say, electron–atom attachment and volume recombination in cases where there is reason to expect that such processes occur.

3.2.2. *The maintained uniform glow.*

It has been assumed that up to breakdown the concentration of electrons and ions is negligible, and this is in accord with experiment. After breakdown, however, the current increases, producing a significant space charge. Initially the electrons diffuse away faster than the ions, so setting up a resultant positive space charge, the radial field of which retards electron diffusion, but assists ion diffusion. In the steady state ions and electrons move away at the same rate to the walls where recombination occurs, but in general no significant radiation is emitted in the process.

A systematic study of high-frequency glow discharges under conditions when diffusion is a controlling factor has been made by Townsend and Donaldson (1928) and Townsend and Nethercot (1929) with cylindrical discharge tubes fitted with external sleeve electrodes. For a fixed gas pressure the field strength X

was constant over the range of frequencies f from 0·5 to 7 Mc/s and discharge currents I from 1 to 12 mA. The parameters satisfied an equation of the form

$$V_m = X(a+x) + bI/f, \qquad \ldots\ldots(15)$$

where V_m is the maintenance potential, a and b are constants depending on the gas pressure and on the diameter of the tube, and x is the electrode separation. The maintaining voltage V_m consisted therefore of two parts: the term $X(a+x)$, which is dependent on the gas pressure only, and the term bI/f, which is dependent on the current and the frequency. The latter term includes the fall of potential between the sleeve and the inner surface of the tube.

A theory of the mechanism of the uniform high-frequency glow has been given by Townsend (1928, 1931, 1932), and the same theory applies to the static field positive column which the high-frequency glow strongly resembles. In long cylindrical tubes the column is electrically and optically uniform along the axis. The longitudinal force X in the uniform glow is independent of the current when heating effects can be neglected. Consequently the number of electrons and positive ions per centimetre are proportional to the current which is constant across any cross section of the discharge, and it may be assumed that the generation of electrons and positive ions is maintained by collision at the rate proportional to the electron concentration. An ionized gas in such a state is sometimes called a plasma and, because of modern applications, the theory of the uniform glow will be given in some detail for a spherical discharge.

The mean electron energy E is proportional to $X/p\sqrt{2}$, where X is the amplitude of the high-frequency electric field. The rate of generation of ions and electrons is $\alpha N_- \omega_- X/\sqrt{2}$ where N_- is the number of electrons in a sphere of radius r. Thus the equations of continuity are

$$\alpha N_- \omega_- X/\sqrt{2} = -4\pi r^2 D_- \operatorname{grad} n_- - 4\pi r^2 n_- \omega_- F(r) \qquad \ldots\ldots(16)$$

$$\alpha N_- \omega_- X/\sqrt{2} = -4\pi r^2 D_+ \operatorname{grad} n_+ + 4\pi r^2 n_+ \omega_+ F(r) \qquad \ldots\ldots(17)$$

where the radial field $F(r)$ is given by

$$F(r) = e(N_+ - N_-)/r^2 \qquad \ldots\ldots(18)$$

and n_- and n_+ are the number of electrons and positive ions per cm^3, ω_-, ω_+ their mobilities and D_-, D_+ their coefficients of diffusion respectively. The radial field is not oscillatory since periodic variations in n_- or n_+ are negligible. The maintaining electric force X increases with the gas pressure, and at high pressures F is small compared with X.

Since n_- is a maximum when $r=0$, dn_-/dr is negative : also D_- and ω_- are large compared with D_+ and ω_+. Consequently it follows that each of the two terms on the right-hand side of (16) is large compared with $\alpha N_- \omega_- X/\sqrt{2}$. Therefore

$$F(r) = -(D_-/n_-\omega_-)dn_-/dr, \qquad \ldots\ldots(19)$$

so that the potential difference between the centre of the glow and the walls of radius a is given by (using (11))

$$-\int_0^a F\,dr = E\log y, \qquad \ldots\ldots(20)$$

where y is the ratio of electron concentrations at the centre to that at the walls, and $D_-/\omega_- = E$. Neglecting $(n_+ - n_-)$ in comparison with n_-, and eliminating $F(r)$ from (16) and (17), it follows that

$$d^2n_-/dr^2 + (2/r)\,dn_-/dr + C^2 n_- = 0, \qquad \ldots\ldots(21)$$

where

$$C^2 = \alpha\omega_- X/\sqrt{2}\,D_a = \alpha\omega_-{}^2 X/\sqrt{2}\,D_-\omega_+ \qquad \ldots\ldots(22)$$

approximately, since $\omega_- \gg \omega_+$ and $D_-\omega_+ > D_+\omega_-$. D_a is the coefficient of ambi-polar diffusion defined by

$$D_a = (D_+\omega_- + D_-\omega_+)/(\omega_- + \omega_+). \qquad \ldots\ldots(23)$$

Putting $n_+ = n_- = n$, the radial wall current j can then be written as

$$j_- = j_+ = -4\pi r^2 D_a \operatorname{grad} n. \qquad \ldots\ldots(24)$$

The solution of (21) is

$$n = n_0 J_0(Cr) \qquad \ldots\ldots(25)$$

and if, following Schottky (1924), n is taken as zero at the walls when $r = a$, $Ca = 2{\cdot}4$. Equation (22) can then be written (Llewellyn Jones 1933)

$$\alpha/p = f(X/p) = A(X/p)^{1/2}/(ap)^2, \qquad \ldots\ldots(26)$$

where A is a numerical factor, giving the similarity relation involving the discharge radius a, so that

$$ap = \phi(X/p), \qquad \ldots\ldots(27)$$

a result which is independent of the energy distribution function. The nature of the ions which reach the walls has been examined for helium discharges by Phelps and Brown (1952) using a mass spectrometer : the ions responsible for the ambi-polar diffusion coefficient at pressures of the order of 5 mm Hg were identified as He^+.

If the current is very small, at breakdown for example, the charge is not sufficient to produce ambi-polar diffusion, and the rate of loss of electrons will then be large. To balance this the ion and electron generation must be increased. This result is in accordance with the fact that the field necessary to start the discharge is greater than that required to maintain the discharge.

The maintenance of microwave discharges in argon has been discussed by Krasik, Alpert and McCoubrey (1949) who determined the (V, I) characteristics for a number of pressures and compared the results with those for the positive column of static field discharges. Llewellyn Jones and Morgan (1951 b) have shown that the extinction potential is not a characteristic of the discharge alone but is dependent on the impedance of the generator. The (V_m, p) curves are similar in form to the (V_s, p) curves which were obtained in air and hydrogen. The curves have a single minimum which decreases with increasing frequency. The minimum maintenance potential of the static field glow discharge is, however, greater than that for the high-frequency glow discharge. The same similarity relationship which applies to breakdown also applies to maintained discharges (Llewellyn Jones and Morgan 1951 b).

3.3. *Electron Energy Distribution*

The value of X/p in the uniform glow of high-frequency (and static field) discharges in helium for example at pressures of the order of 1 mm Hg in wide tubes are about 1·5 to 3 v/cm/mm Hg (Llewellyn Jones 1931). Now Townsend (1924) has shown that the mean electron energy E is about $2X/p$ in helium, so that

in the uniform glow $E \simeq 3$ to 5 ev in both static and high-frequency fields. It is clear, however, that since excitation and ionization of neutral helium atoms require electron energies exceeding about 22 ev, the electron energies must be distributed about a mean value. The form which the distribution function takes is of paramount importance in the quantitative understanding of many optical and electrical properties of discharges. The relative intensities of the spectral lines and the variation of intensity with gas pressure are important effects which are accounted for by the form of the electron energy distribution function. Since X/p diminishes (slowly) as p increases, a change of gas pressure consequently involves a small change in the mean electron energy, and the resulting change in the number of electrons with energies exceeding the resonance energy of the gas atoms depends on the form of the distribution function.

In weak static electric fields, when the electrons are nearly in thermal equilibrium with the gas molecules, the electron energy distribution is nearly Maxwellian, but as the field is increased the mean electron energy is increased, and the distribution changes to one which is determined by diffusion, drift and energy losses in collisions. Theoretical formulae have been deduced by Townsend, Druyvesteyn, and by Smit for static fields (Loeb 1939, Druyvesteyn and Penning 1940), and by Holstein (1946 b) and Margenau (1946, 1948 b) for high-frequency fields: for certain values of X/p and f the distribution is the same in both cases. For an infinite gaseous space the motion of the electrons when the frequency of the field is lower than the collisional frequency of the electrons in the gas is almost the same as under static field conditions. In an ideal gas the scattering of the electrons will be uniform, and the direction of the field is of no significance. The current due to the drift of electrons will be in phase with the field and the energy which is transferred will be the same as for a static field. Properties shown experimentally to be the same in both high-frequency and static discharges in the same gas and tube are the mean electric field (Townsend 1930, 1933, Johnson 1930, Llewellyn Jones 1931, 1933) and the variation of spectral line intensities with pressure (Townsend and Llewellyn Jones 1931 a, b). Under these conditions X/p is the same, and it follows that α/p, the diffusion coefficient, the drift velocity and electron energy distribution will be the same in the two kinds of discharges.

From the measured variation of the intensities of the helium lines (Keyston 1933) Llewellyn Jones (1936, 1944) deduced the electron distribution functions in that gas. The form varies somewhat with X/p, and for mean values of the order of 4 ev was broader than the Druyvesteyn distribution for an ideal gas, but was narrower than the Maxwellian form except at the higher electron energies. This is accounted for by the form of the variation of the atomic cross section for elastic scattering of electrons, which in helium tends to increase the number of electrons in the higher energy group above that corresponding to the Druyvesteyn form. The form of the variation of atomic cross sections also accounts for the extremely narrow form of distribution in argon at high mean energies ($\sim$8 ev) found from a study of line and continuous spectra by MacCallum, Klatzow and Keyston (1933). For low mean energies ($\sim$0·5 ev) Huxley and Zaazou (1949), using the Townsend diffusion method, showed that the electron distribution function in air approximates to the Druyvesteyn form, as expected in theory. Thus care must be exercised in applying the formula for the distribution function calculated for an ideal gas: the Townsend–Ramsauer effect plays an important part in determining the distribution, which varies over very wide limits from one gas to another. For example, the

distribution for low mean energies can be very narrow in argon, practically the Druyvesteyn ideal form for neon, and is rather wider than Maxwellian in helium. Expressions for α can be evaluated when the distribution and excitation functions are known (Abdelnabi and Massey 1953).

At very high frequencies the efficiency of the transfer of energy to the electrons decreases (Margenau 1946, 1948 b).

3.4. *Applications*

3.4.1. *Atomic collisional phenomena.*

Recent advances in microwave techniques have enabled the coefficients of recombination and diffusion to be calculated (Biondi and Brown 1949 a, b) from the measurement of the change Δf of the resonant frequency f of a microwave cavity due to the presence of an electron concentration n_- in the cavity (Slater 1946). Under certain conditions Δf is linearly proportional to the electron density within the cavity

$$\Delta f = f - f_0 = Cn_- \qquad \ldots\ldots(28)$$

where f_0 is the resonant frequency of the cavity in the absence of electrons and C a geometrical coefficient determined by the distribution of electron density and electric field within the cavity. It is assumed in the derivation of (28) that the presence of electrons perturbs the cavity only slightly (Brown and Rose 1952). The principle of the method of measuring the coefficients of ambi-polar diffusion, from which is derived the mobility of ions, is as follows:

In a decaying glowing plasma in which there is no source of ionization $\mathrm{div}\,\boldsymbol{j} + \partial n_-/\partial t = 0$, so that

$$-D_a \nabla^2 n_- + \partial n_-/\partial t = 0. \qquad \ldots\ldots(29)$$

If n_- decreases exponentially with a mean decay time τ, then $n_- = n_0 \exp(-t/\tau)$, where n_0 is the concentration at $t = 0$ when the ionizing agency is cut off. Thus from (21) n_- is given by

$$n_- = B\{\sin(1/D_a\tau)^{1/2} r\}/r, \qquad \ldots\ldots(30)$$

where $$(1/D_a\tau)^{1/2} = R = S\pi \text{ (where } S = 1, 2 \ldots, n_- = 0 \text{ at } r = a). \qquad \ldots\ldots(31)$$

B is a constant and R is the radius of the spherical container. The allowed values of τ correspond to different modes of diffusion, and in general the concentration n_- will be a linear combination of different modes determined by the distribution when the ionization is cut off. If when the ionization is cut off the distribution is that corresponding to the lowest mode (i.e. when $S = 1$) then

$$\tau = R^2/\pi^2 D_a. \qquad \ldots\ldots(32)$$

Therefore if τ is measured then D_a can be calculated.

If the decay of electron density is observed after the electrons and ions have reached equilibrium with the gas, and so possess a Maxwellian distribution at a temperature T, then since $\omega_- \gg \omega_+$

$$D_+/\omega_+ = D_-/\omega_- = \boldsymbol{k}T/\boldsymbol{e} \qquad \ldots\ldots(33)$$

and
$$D_a = 2D_+\omega_-/(\omega_+ + \omega_-) \sim 2D_+ \qquad \ldots\ldots(34)$$

or
$$\omega_+ \simeq D_a/2\boldsymbol{k}T. \qquad \ldots\ldots(35)$$

In the experiments of Biondi and Brown helium was ionized by a 250 μsec pulse from a magnetron ($\lambda = 10$ cm). During the pulse a stationary charge distribution was established of concentration approximately 10^{11} electrons or ions per cm^3. With the discharge tube in the uniform field the charge distribution

corresponded to the lowest mode of diffusion ($S=1$). The magnetron was then turned off for 11 msec, and during this period the change of electron density with time was measured by noting the change in the resonant frequency of the cavity. Only a period of 50 μsec was necessary for the establishment of thermal equilibrium of the ions, electrons and gas molecules.

The rate of decrease of n_- at pressures about 5 mm Hg, when both diffusion and some recombination (of coefficient η) act, is given by

$$n_-/(1+\eta\tau n_-)=\{n_0/(1+\eta\tau n_0)\}\exp(-t/\tau) \qquad \ldots\ldots(36)$$

where n_0 is the electron concentration at $t=0$, which was taken 3 msec after ionization ceased. Since η is independent of pressure this relation may be used to correct the data for any recombination loss. At higher pressures ($\sim$20 mm Hg), however, when volume recombination predominates and diffusion may be neglected, (36) becomes

$$1/n_- - 1/n_0 = \eta t. \qquad \ldots\ldots(37)$$

Biondi and Brown found for such cases that the curves relating $1/n_-$ with t were linear, and concluded from this that recombination of ions and electrons was taking place. After the discharge was terminated, an after-glow, which lasted for more than 4000 μsec, was observed, and this was taken as evidence supporting the view that recombination was taking place.

Biondi and Brown in this way measured η and ω_+ in the noble gases, hydrogen, nitrogen and oxygen. At certain temperatures η could be expressed as the sum of two coefficients, one dependent on p and another independent of p. The values of η obtained by Biondi and Brown in helium ($\sim 10^{-7}$) are larger than those ($\sim 10^{-9}$) later found by Johnson, McClure and Holt (1950) with the same method, but in neon Biondi and Brown's value ($\sim 10^{-7}$) practically agreed with the later determinations of Holt, Richardson, Howland and McClure (1950). In diatomic gases Biondi and Brown (1949 b) showed that η decreased with atomic number, and in hydrogen was independent of pressure at 300°K, maintaining a value of $2{\cdot}5\times 10^{-7}$ cm^3/sec over the pressure range below 13 mm Hg. Thus the coefficient η differs greatly for different gases ; furthermore, values of η for different samples of the same gas which differ in purity differ by a large factor. Values obtained by the microwave technique are within the limits 10^{-6} to 10^{-9} cm^3/sec, and these are several orders of magnitude greater than both the earlier determinations (10^{-10}) by Kenty (1928) and the values calculated for radiative recombination (10^{-12}) by Stueckelberg and Morse (1930). These results are discussed by Massey (1952) and conclusions are reached about the nature of the ions and the reactions involved in the recombination process.

Recombination coefficients of ions with ions for air, however, have previously been measured and discussed by Sayers (1938, 1943, 1947, 1950), and further work on the recombination of electrons with ions and of ions with ions is in progress. The effects of recombination and attachment on the ionization coefficient α have been studied by Geballe and Harrison (1952).

Phelps and Brown (1952) have shown that while the values of the ambi-polar diffusion coefficient in helium obtained by Biondi and Brown (1949 a) related to the atomic ion He^+, the recombination coefficient obtained at higher pressures related to the molecular ion He_2^+. Bates (1950) attributed the high values of η obtained by Biondi and Brown to the dissociative recombination of He_2^+. Phelps and Brown (1952) have also measured D_a in helium and found a value

of 14 cm^2/volt/sec for the mobility of He^+ in He. With regard to these estimates of ion mobilities, it is relevant to note that Hornbeck (1951), using a transit time method, has measured simultaneously the mobilities of both He_2^+ and He^+, obtaining values of 19 and 10·8 cm^2/volt/sec respectively ; the theoretical value calculated for He^+ in He is 12 cm^2/volt/sec (Massey and Mohr 1934). It had earlier been suggested by Meyerott (1944) that Tyndall and Powell (1931) had observed the mobility of He_2^+ ions and not He^+ ions. Hornbeck (1951) measured ω_+ for Ne_2^+ and A_2^+, and the values were the same as those ascribed by Munson and Tyndall (1941) to Ne^+ and A^+. A mass spectrometer probe for discriminating between ions of atomic and molecular mass by a time of flight method has been described by Boyd (1950a). Preliminary experiments (Boyd 1950b) indicate that values of between 2 and 5 may be obtained for the ratio of He_2^+ to He^+ in a hot-cathode helium discharge at various pressures and currents.

Studies of atomic cross sections for slow electrons in different gases have been made, using microwave techniques, by Phelps, Fundingsland and Brown (1951). Preliminary measurements of the collision probability P_e for electrons with mean energies of 0·039 ev at 1 mm Hg in units of cm^2/cm^3 are He 19, Ne 3·3, A 2·1, Kr 54, Xe 180, H_2 46, N_2 15. The values of P_e in He, Ne and A are in good agreement with the theoretical calculations of Allis and Morse (1931). Similar measurements have also been made by Varnerin (1951). The importance of the microwave method is that data may be obtained when the electron and gas atoms are in thermal equilibrium. In addition, the frequency of the applied field can be made much greater than the average collision frequency of the electron at pressures such that the mean free path of the electron is small compared with the dimensions of the discharge tube.

3.4.2. *Sources* (*of ions and radiations*).

The yield of high-frequency ion sources, of the type often used in particle accelerators, is high because the absence of electrodes reduces recombination and the low electric fields ensure that the spread in the energy of the emergent ion beam is small. Such sources have been widely investigated (Rutherglen and Cole 1947, Neuert 1949, Thonemann, Moffatt, Roaf and Sanders 1948, Moak, Reese and Good 1951).

Electrodeless discharges are often used with advantage as spectroscopic sources (Nisewanger, Holmes and Weissler 1946) since the absence of internal electrodes eliminates a source of impurity and the low electric fields reduce the Stark broadening.

The study of high-frequency discharges is of great importance in the design of radio-frequency switches, and the application of high-frequency torches for welding refractory materials is of great practical interest (Cobine and Wilbur 1951). High-frequency discharges can also be used as detectors of ionizing radiations (Brown and McCarthy 1948), and it is claimed that this type of detector is faster by one or two orders of magnitude than the conventional Geiger counter, in which a static field is applied to remove the residual ions.

3.4.3. *Ionospheric non-linear properties.*

An interesting application of the fundamental principles described above in §3.3 concerning the build-up of the mean electron energy in high-frequency fields lies in the study of the propagation and interaction of electromagnetic waves in the

ionosphere. The conditions differ from those in a low-pressure discharge tube because of the absence of wall effects. As in a discharge tube, however, the electrons and gas molecules are in thermal equilibrium in the absence of an electric field. When a high-frequency field is applied, the electrons acquire energy and the mean electron energy rises until the rate of gain of energy from the field is equal to the rate of loss in collisions with gas molecules. Consequently the effect of an applied field is to increase the collision frequency and so to decrease the reflecting power of the ionosphere. If the applied field is modulated, this modulation will be impressed upon any radio wave reflected from the part of the ionosphere which is affected : in this way non-linear properties are introduced. This phenomenon, which is now known as radio wave interaction, was first reported by Butt (1933) and Tellegen (1933). An explanation in the above terms was proposed by Bailey and Martyn (1934) and later revised and extended to include gyro-interaction by Bailey (1937). Radio interaction has recently been more intensively studied and used as a tool for investigating the ionosphere and estimating the electron collision frequency and density gradient (Huxley, Foster and Newton 1948, Ratcliffe and Shaw 1948, Huxley 1950, Shaw 1951). Although the existence of gyro-resonance reaction has been questioned, the phenomenon has been reported by Cutolo (1947, 1950), and applied to measure the intensity of the earth's magnetic field in the ionosphere. More recent experiments (Bailey, Smith, Landecker, Higgs and Hibberd 1952) further confirm Bailey's theory, which, in an extended form, allows estimates to be made of the magnetic field, gas temperature, electron collision frequency, density gradient, and the collision energy loss coefficient.

§4. COLD ELECTRON EMISSION IN DISCHARGES

4.1. *Occurrence*

The liberation of electrons from cold metal surfaces *in vacuo* by the application of intense electric fields has been reviewed by Jenkins in *Reports on Progress in Physics*, Vol. IX, 1942–43, to which reference should be made. This section is concerned with the cold liberation of electrons from electrodes in gases, the mechanism of the process, and its role in electrical discharges.

Before the development of the wave-mechanical theory of field emission by Fowler and Nordheim (1928) it had been suggested that electrons drawn out of the cathode of a discharge gap by high fields played an important part in the maintenance of an arc current. Langmuir (1923), to explain the mechanism of the mercury-arc discharge, supposed that electrons were liberated from the cathode by a high electric field set up by a space charge layer of positive ions in the discharge. Mackeown (1929), Compton (1931), Ramberg (1932), Tonks (1934, 1935), Becken and Sommermeyer (1936) and Seeliger and Becken (1935) have all put forward the view that field emission provides the necessary high current density in arcs from low boiling point metals. In their review Druyvesteyn and Penning (1940) suggested that because the cathode spot of an arc can in many cases be made to move rapidly, the emission cannot be thermionic, and that field emission seemed to be the most probable explanation. Langmuir (1928), to explain the failure of mercury-arc rectifiers during the reverse phase (often termed back-firing), postulated field emission caused by the presence of insulating material which is highly charged by positive ions. This was also

suggested by Druyvesteyn (1936) in work on arcs. The suggestion that field emission from the cathode maintains cold arcs leads, however, to difficulties because the electric fields existing at the cathode ($\sim 10^6$ v/cm) are too low to give quantitative agreement with the Fowler–Nordheim theory. This lack of agreement appears to be a feature of all work on arcs because, in order to maintain by field emission the current densities known to occur, the electric field should be greater than about 10^7 v/cm.

The suggestion that field emission can account for the sudden transition from a glow discharge with a comparatively low current density to a high current density arc discharge has often been made. In a comprehensive investigation of glow-to-arc transition phenomena in air, oxygen, nitrogen and hydrogen with water-cooled electrodes of Cu, Al, Mo and Pt, Fan (1939) concluded that the cathode surfaces could not reach the temperature required to give appreciable thermionic emission : he calculated the field strength in the discharge and showed that intensities of between 10^5 and 10^6 v/cm could exist at the cathode surface, and concluded that field emission could arise on account of the local concentration of the electric field by small irregularities at that surface. A similar suggestion was made by Jurriaanse and Druyvesteyn (1936), and the work of Wehrli (1927) in this connection is of interest.

A special form of glow discharge characterized by the absence of dark space and cathode fall, and by a heavy current even at low voltage and gas pressure, has been described by Guntherschulze and Fricke (1933). This was further investigated by Guntherschulze and Bär (1937), Schnitger (1935, 1936) and Wolf (1939). The cathodes used were non-metallic, usually carbon or ocelite, and covered by a layer of an insulating powder, e.g. aluminium oxide particles or electrolytically formed aluminium oxide were usually employed. The relation between the current density J and the potential difference V between the cathode and anode was found by Schnitger to be

$$J = AV^3 \exp(-B/V) \qquad \ldots\ldots(38)$$

where A and B are constants. This shows that the mechanism of electron liberation was a field-dependent process. The equation is of course of the same analytic form as the Fowler–Nordheim field emission equation. The action of positive ions in setting up an intense field across the aluminium oxide was regarded as the cause of the emission: the high field produced electrons with energies approximately equal to the potential difference across the gap, and these could ionize immediately. Schnitger concluded from these results that it was essential that the material under the insulating layer should have a high resistance. The average value of the field set up by positive ions was estimated to be about 5×10^5 v/cm ; but this value does not give quantitative agreement with the Fowler–Nordheim equation.

As indicated above (§2.3.1), the use of gases at high pressure as an insulating medium in power cables, transformers, transmitter condensers, air-blast circuit breakers, high-voltage generators used for nuclear research, etc., has led to a large number of investigations of the static sparking potentials of gases at high pressures p. These investigations have usually been carried out for comparatively small gap distances and hence at high electric fields X. The work of Zeier (1932), Howell (1939), Skilling (1939), Skilling and Brenner (1941), Trump, Safford and Cloud (1941), Felici and Marchal (1948), Young (1950) and Trump, Cloud, Mann and Hanson (1950) has shown that a pronounced failure of Paschen's law occurs. The

investigators agree that this departure was due to the significant occurrence of a source of ionization which did not depend on the discharge parameter X/p, being instead a function of the field X and the nature of the cathode surface. It follows that the electron emission from the cathode must play an important part in the breakdown mechanism, which then cannot be accounted for in terms of the gas processes alone. The results of Skilling and Brenner and those of Trump *et al.* indicated that the electron emission became the controlling secondary ionization process in the breakdown mechanism at the highest pressures investigated. Although all observers have suggested that the liberation mechanism was cold field emission from the cathode at high values of X, which in these experiments was of the order of 10^5 v/cm, no measurements of the magnitude of the field current under such circumstances have been given. It is, however, clear that the current should be large, and this again leads to quantitative disagreement with the emission estimated from the Fowler–Nordheim equation. In spark time lag studies Cobine and Easton (1943) and Llewellyn Jones (1946, 1949 a) suggested that field emission from oxidized cathodes in air and in nitrogen at fields as low as 10^5 v/cm was responsible for the very rapid initiation of sparks. They envisaged a mechanism similar to that suggested by Langmuir (1928), which was investigated by Paetow (1939). No quantitative measurements of the magnitude of the field emission currents in their experiments were given.

In all the cases given above there is an apparent discrepancy between the magnitude of the average electric field obtaining in the experiments ($\sim 10^5$ to 10^6 v/cm) and that at which the field emission current should become significant on the basis of the Fowler–Nordheim equation ($\sim 10^7$ to 10^8 v/cm), and this discrepancy requires investigation.

4.2. *Investigation of the Mechanism of Emission*

To establish whether the field emission of electrons can contribute to the discharge mechanism it is necessary to determine the quantitative dependence of cold electron emission upon the electric field. Such a study has been carried out in recent years (Llewellyn Jones, de la Perrelle and Morgan 1951). The technique employed was an extension of that previously employed by Llewellyn Jones (1949 a), and consisted of the measurement of the mean statistical time lag of spark breakdown of a short spark gap of which the cathode was the metal under investigation. Under certain conditions (Llewellyn Jones and de la Perrelle 1951, 1953) the mean statistical time lag bears a simple relation to the rate of pre-breakdown cathode emission in a discharge gap.

The principle of the method was as follows. A series of rectangular pulses of e.m.f. was applied to the gap in which pre-breakdown ionization due to natural processes was zero, and the time which elapsed before the appearance of each electron after the sudden application of the e.m.f. was measured electronically. The mean value $\bar{t}$ of about 1000 of these times was then determined, and this gives the rate of electron emission I from the cathode under the impulses of e.m.f. according to the relation

$$\bar{t} = 1/I, \qquad \ldots\ldots(39)$$

and the distribution of the time lags about this mean value is given by $I \exp(-t/t)$. It follows that the number of lags greater than or equal to t out of a set of N lags is given by the relation

$$\ln(N/n) = It, \qquad \ldots\ldots(40)$$

so that I could also be obtained from a graph of $(\ln N/n, t)$. This permitted two independent estimates of the I to be made. In these experiments great care was taken to ensure that the necessary conditions imposed by the theory were satisfied. In previous investigations (Zuber 1925, Tilles 1934) all these conditions were not satisfied: curvature of $(\ln N/n, t)$ graphs can result from failure to fulfil the necessary conditions, making it impossible to find I.

Electron emission from nickel and tungsten electrodes in various surface states was measured, and it was shown that considerable emission rates of the order of 10^5–10^6 electrons/sec could be obtained when the applied electric field was as low as 10^5 v/cm. This current was highly dependent upon the nature of the surface, its previous history, and the presence of contamination such as dust particles, oxidation, etc. With highly polished cathodes the emission was found to decrease rapidly from an initial value of approximately 10^6 electrons/sec to zero when the cathode was subjected to low current sparking. Such a decrease in electrode activity had previously been noted by Howell (1939) and others, and is usually referred to as *conditioning* or *formation*. It is attributed to the removal of dust particles and other sub-microscopic surface irregularities by erosion in the spark discharges of very low currents. The fact that surfaces from which the emission was zero could be obtained by subjecting highly polished surfaces to low-current sparking was important in the development of the experimental technique.

With heavily oxidized dusty electrodes the cold emission was greater than 10^6 electrons/sec, and removal of the dust produced a reduction of about 10^3 times in the emission. Thus with heavily oxidized electrodes an enhanced emission of the order of 10^4–10^5 electrons/sec was obtained with an applied electric field of 2×10^5 v/cm.

The source and mechanism of the enhanced emission obtained from nickel and tungsten cathodes covered with an oxide layer was investigated by measuring time lags for different values of the electric field X (Llewellyn Jones and de la Perrelle 1953). A typical set of curves for the nickel electrode is given in fig. 12, in which n refers to the number of lags greater than time t of a set of N measurements.

The slope of the $(\ln N/n, t)$ graphs gives the electron emission I in electrons/sec corresponding to the various values of X. By plotting $\log I/X^2$ against $1/X$ a straight line is obtained as given in fig. 13, which also gives the relations found with tungsten. This shows that the relationship between I and X is

$$I = A X^2 \exp(-D/X), \qquad \ldots\ldots(41)$$

where A is the intercept on the $\log I/X^2$ axis and D is the slope of the straight line in fig. 13.

It is of interest to compare the above experimentally established equation with the Fowler–Nordheim (1928) equation for the field emission of electrons from a solid surface under a field F:

$$I = 38{\cdot}5 \times 10^{12} S F^2 \frac{\zeta^{1/2}}{(\zeta+\phi)\phi^{1/2}} \exp\left(-6{\cdot}8 \times 10^7 \frac{\phi^{3/2}}{F}\right) \qquad \ldots\ldots(42)$$

where I is the number of electrons emitted per second from a surface of area S, ϕ is the work function, and ζ the Fermi energy in volts of the solid. For a perfectly smooth surface the field F would be the same as the average field X in the gap; in practice, however, $F = MX$ where M is a factor greater than 1, which allows for the local intensification of the field due to surface irregularities.

Comparing eqns. (41) and (42), it may be seen that the experimental results allow estimates of ϕ and S to be made provided that a numerical value can be given

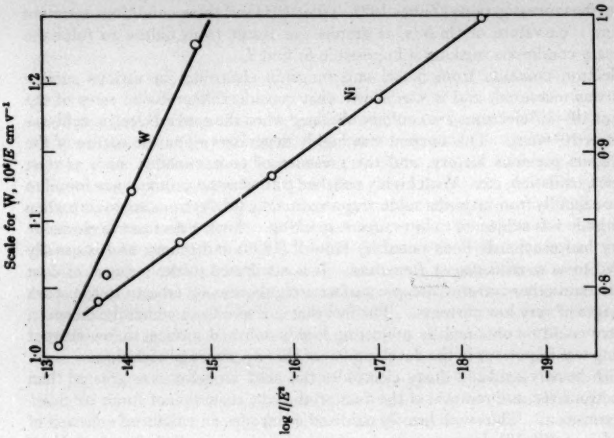


Fig. 13. Field emission from nickel and from tungsten.
(Reproduced by courtesy of the Royal Society)

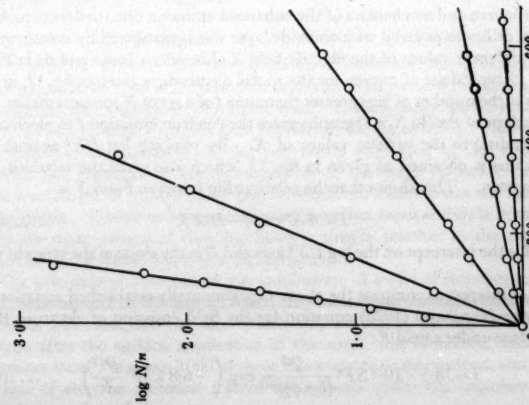


Fig. 12. Log (N/n), t curves obtained with nickel cathode.
(Reproduced by courtesy of the Royal Society)

to *M.* Schottky (1923) has shown that it is unlikely that surface irregularities could enhance the field by more than a factor of about 10. From calculations on similar lines, and also optical examination of the surfaces employed, it seemed

improbable that M greatly exceeded 10. Estimates of ϕ and S and their dependence on the value chosen for M are given in the table.

M	1	10	100	1000
ϕ (ev)	0·1	0·48	2·3	10·8
S (cm^2)	10^{-13}	2×10^{-14}	6×10^{-15}	3×10^{-17}

It is seen that when M lies between 1 and 10, S is of the order of the atomic cross section ; on the other hand, the higher values of M ($\sim10^2$ to 10^3) lead to estimates of S and ϕ of less than 10^{-16} cm^2 and about 11 ev respectively. Such values can have little physical significance. It appears therefore that M must lie between 1 and 10 in order to obtain reasonable values of both ϕ and S together from the Fowler–Nordheim equation.

The question raised by consideration of the values of both S and ϕ calculated from the Fowler–Nordheim equation is very important. It appears that in previous work on field emission *in vacuo* insufficient attention has been paid to the calculations of S, most emphasis having been paid to the value of ϕ obtained. This is significant. In previous experiments on field emission (see Jenkins 1943 for references) there has apparently always been a discrepancy between the emitting areas calculated from the Fowler–Nordheim relation and estimates of this area made from the geometry and dimensions of the cathode. This lack of agreement between theory and experiment, which in some cases can involve a factor of the order of 10^2 in the emitting areas, is consistent with an interpretation in terms of electro-positive contamination on the cathodes (Stern, Gossling and Fowler 1928). However, in the present case of thick surface layers, and in many other of the investigations referred to by Jenkins (1943), it is unlikely that such an explanation is valid.

It seems reasonable to suppose then that the sources of initiatory electrons in these cases are possibly groups of negative ions on the oxide layer at the cathode surface. The estimates of the emitting area when $M\simeq10$ are of molecular dimensions and the effective work functions of the order of 0·5 ev. These values are in accordance with this view, since oxygen negative ions have a detachment energy of about 0·4 ev (Massey 1950). These results show that the occurrence of high cold emission under comparatively low fields was not limited to the case of an insulating layer on a high resistance base, as concluded by Schnitger : the emission takes place from oxide layers on metals.

4.3. *Role of Thin Surface Films*

Using this same technique, another class of surfaces was investigated by Llewellyn Jones and Morgan (1953). These surfaces are apparently clean, but are actually covered with a thin tarnish film about 10^{-7} cm thick. Such films on nickel, iron, aluminium and aluminium alloy were examined in dry air, hydrogen and argon. In each case considerable field dependent emission I of the order of 10^5 electrons/sec was obtained with applied electric fields X as low as 10^4 v/cm ; the relation between I and X was again of the form (41). Comparison with the Fowler–Nordheim equation, however, leads to estimates of S and ϕ of 10^{-20} cm^2 and 10^{-2} ev respectively. Since these small values can have no physical significance it is clear that the simple Fowler–Nordheim theory does not describe the emission in these cases, and some other mechanism must be sought.

A feature common to all these experiments in dry gases, in which the electrodes were covered with thin tarnish films, was that a long delay (in some cases as much as two minutes) was noted between the application of the first voltage impulse to the gap and the occurrence of the first spark, i.e. the liberation of the first electron. All subsequent applications of the voltage impulses in a given set led to sparks with lags distributed about a very small mean value ($\sim 10^{-3}$ to 10^{-5} sec). Clearly some process initiated by the occurrence of the first spark played a predominant role in liberating electrons which led to all the subsequent sparks, the first spark being initiated by an electron liberated by natural processes.

A possible explanation is as follows (Llewellyn Jones and Morgan 1953). Positive ions produced in the first spark eventually reach the cathode and then set up an intense electric field across the thin surface tarnish film. The potential energy diagram in this case is similar to that shown in fig. 14, where F_1 is the intense micro-field set up by the positive ions across the tarnish film, of thickness

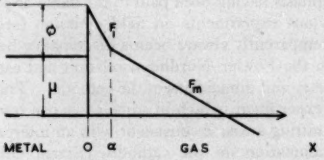


Figure 14. Potential diagram for cathode surface covered with thin film.
Note. μ in this diagram corresponds to ζ in the text.

a, and $F_m(=X)$ is the applied electric field. Now this potential barrier is of exactly the same form as that treated by Stern, Gossling and Fowler (1928) for the emission of electrons through a thin film of sodium on tungsten. In their case the field F_1 at the cathode was considered to be due to the contact potential difference between the sodium and tungsten and given by

$$F_1=(\phi_1-\phi_2)/a. \qquad \ldots\ldots(43)$$

Stern, Gossling and Fowler showed that in this case the current I was given by

$$I=38{\cdot}5\times10^{12}\frac{\zeta^{1/2}SF_2^{\,2}}{(\zeta+\phi_1)^{1/2}\phi_1^{\,1/2}}\exp\left(\frac{-6{\cdot}8\times10^7\phi_2^{\,3/2}}{F_2}\right)$$
$$\times\frac{\phi_1}{\phi_2}\exp\left[-6{\cdot}8\times10^7 a\left\{\frac{\phi_1+\phi_2+(\phi_1\phi_2)^{1/2}}{\phi_1^{\,1/2}+\phi_2^{\,1/2}}\right\}\right] \ldots\ldots(44)$$

where $F_2=MX$. This modified equation may be applied to the present case. In so doing the slope of the $(\ln I/X^2, 1/X)$ curve which, on the Stern, Gossling and Fowler theory, gives ϕ_2, may be used to estimate F_1 from the above relation. The equation also permits estimates of the emitting area to be made, but it is necessary to know the film thickness a. For aluminium, for example, taking $a=2\times10^{-7}$ cm (Cabrera and Mott 1949), the values of S became 10^{-7} cm^2 to 10^{-13} cm^2 instead of 10^{-20} cm^2 on the simple Fowler–Nordheim theory.

These are the areas from which the emission took place, i.e. on the present view they are the areas on which the positive ions were deposited. As these ions were produced in a previous spark the areas should be about the same order as the area of

these spark channels at the cathode. The areas ($\sim 10^{-6}$ cm^2) of such spark cross sections at the cathode have been calculated by Llewellyn Jones (1950), and these values agree with the experimental estimates of Froome (1948).

A test of the role of the film was made by gradually removing it by hydrogen ion bombardment and measuring the cold emission at stages during the removal: the emission decreased and practically vanished. In this way it was established that the film was responsible for the emission, while other experiments showed the necessity for the presence of positive ions.

To sum up, it appears then that the mechanism of electron emission from electrodes in discharges depends upon the thickness of the surface oxide film. For relatively thick surface films of about 10^{-5} cm or more the mechanism is a pure field emission process and the source of electrons appears to be a collection of negative ions which are always present on apparatus exposed to air. For thin films of about 10^{-7} cm or less, on the other hand, the electrons are liberated from the base metal under the influence of the high field set up by positive ions on the film: the positive ions play a dominant role.

It is now of interest to consider whether such field processes can influence the secondary ionization mechanism in such a way as to produce a modification of Paschen's law (see §2.3.1). Trump *et al.* (1950) have suggested that the high fields employed in small gaps at high gas pressures may enhance the secondary ionization coefficient γ, and that such enhancement would lead to a lowering of the sparking potential V_s and to failure of Paschen's law. It has been established that considerable field emission is obtainable from oxide layers on a cathode surface, and it is interesting to consider whether the presence of positive ions on such layers could produce an enhancement of the microscopic field there, and thus produce increased field emission in accordance with the equation above. The net result of this would be a comparatively high electron emission due to the incidence of the positive ions on the cathode or, in other words, an enhancement of the effective value of γ as compared with the (very low) values found in low pressure work when the field X is low. Previous results have shown how rapidly the electron emission I increases with the field X, so that an enhancement of X due to the presence of positive ions on cathode surface layers could lead to greatly increased electron emission. Such a mechanism would take place during the whole pre-breakdown process and would appear as a greatly enhanced γ.

It is not, however, unreasonable to consider that the efficiency of the effective γ process would be progressively increased by very large factors. This would be especially the case when there was considerable pre-breakdown concentration of positive ions in the gap during the time when the sparking potential was being measured. It is significant to note that cold electron emission of the order of 10^4 times greater than would be expected for clean surfaces has been obtained with macroscopic fields of approximately 10^4 v/cm in the presence of oxide films, so that an effective enhancement of this order of the coefficient γ due to the presence of positive ions of the gas on this oxide layer on the cathode would appear to be possible (Llewellyn Jones and Morgan 1951).

The enhanced emission at low fields leads in many cases to a high probability of discharge initiation and is possibly the cause of back-fires in mercury-arc rectifiers, and probably initiates arcing in low- and high-power contact devices and can be responsible for the incipient breakdown of high-voltage equipment in general. There are also some other applications of this phenomenon of

emission from metals covered with tarnish films. The thin film mechanism can play a dominant part in the development of vacuum breakdown, for in this case positive ions may be produced by the impact of high-energy electrons at the anode and by the ionization of adsorbed gases emitted from the electrodes.

§5. THE REGIME OF SPACE CHARGE—THE DISCHARGE ESTABLISHED

5.1. *D.C. Arc Discharges*

After breakdown occurs in a gap, the current can increase almost indefinitely as far as the gas processes are concerned ; in practice, however, the discharge is limited by the external circuit. As the current increases the positive ion space charge density also increases, and this increase is greatest at the cathode where the field in then concentrated. This is the region where, in general, the ionization occurs which produces the initial electrons necessary for the maintenance of the discharge. It is generally considered that this current density J in this important regime is controlled by the Langmuir–Childs space charge equation

$$J = 3{\cdot}85 \times 10^{-9}\, V^{3/2}/d^2 \text{ A/cm}^2. \qquad \ldots\ldots(45)$$

It follows from the similarity relationships (27) that the radius of the region of ionization, i.e. the discharge radius, falls as p increases for a given current when X/p is either constant or diminishes. Thus at high pressures the discharge becomes constricted and the current density consequently increases according to the relation

$$\phi(J/p^2) = \text{const.} \qquad \ldots\ldots(46)$$

The interchange of energy between electrons, ions and the gas molecules then increases with the result that energy is transferred from the field to the gas molecules. The gas temperature consequently rises, and very high gas temperatures can in this way be attained. The temperature thus becomes an important discharge parameter, controlling the additional loss of energy from the arc current by the radiation and conduction of heat from the discharge surface or from the walls of the containing vessel.

In the arc discharge there are the usual dark spaces near the electrodes, but if the arc is long enough, a bright hot uniform glow is maintained : this is the positive column or plasma. Although the glowing gas is itself hot, experience shows that the cathode can in some cases be hot but in others cool (Druyvesteyn and Penning 1940). Clearly the cathode plays a basic role in supplying the initial electrons which are necessary to maintain the d.c. arc, and one of the fundamental problems of the arc discharge is concerned with elucidating the mechanism of this electron production. In general it is considered that the hot cathode arc produces electrons by thermionic emission, while in the case of the cold cathode arc the mechanism must be some process dependent on the electric intensity at the surface, but there is still considerable difficulty in accounting quantitatively for all experimental results (see §4).

These two aspects of the arc problem—the electrode phenomena and the gas processes—have received considerable attention in recent years. The processes of the transfer of ions and electrons through the glowing gas across the gap represent the mechanisms by which the initiatory electrons with those they create in the gas are lost and their number replaced. The arc discharge is of considerable interest owing to its widespread occurrence in electrical engineering; and the suppression

of arcs of various forms is an important problem in electrical contacts of all kinds, from the high power circuit breaker to the low voltage telephone relay. Developments in the electrolytic extraction of metals and demands for stringent speed control of rolling-mill motors have led to a large increase in the use of mercury-arc rectifiers, and, again, the development of the high-pressure mercury arc lamp for medicinal and lighting purposes has focused attention on the thermal and optical properties of the arc. Extensive use has been made of the fact that ions may be extracted from the arc column by the application of a sufficiently strong magnetic field, and a number of ion sources based on this principle are in use. The particular case of the high-pressure mercury vapour discharge and the application of the theory to the design of discharge lamps has been comprehensively dealt with by Elenbaas (1951), who has considered in detail the mechanism of the glowing column and the influence of small traces of cadmium and rare gases. The use of arcs as ion sources received considerable stimulus under wartime conditions by the demand for the separation of uranium isotopes, and under the aegis of the Manhattan Project a considerable amount of theoretical and experimental work was carried out. In this work high current densities were used by constricting the discharge with an applied magnetic field. The collected reports have been published in one volume (Guthrie and Wakerling 1949), to which reference should be made.

5.1.1. *Theory of the arc column.*

Various theories have been presented to describe the electronic and ionic processes, the particle energy distribution and concentrations, and the field distribution in the arc column in terms of the fundamental collisional processes. In all these theories it is generally assumed that since the temperature is high there is almost complete equipartition of energy between the various charged particles, which then have a Maxwellian distribution at the thermal temperature of the glowing gas. Since, however, the ions and electrons still gain energy from the electric field, even at these high temperatures, this assumption can only be an approximation.

Tonks and Langmuir (1929 a, b) proposed a theory of the positive column at low pressures and deduced five fundamental relations which are necessary and sufficient to describe the low-pressure plasma. These five relations are a development of the theory of the positive column put forward by Schottky (1924) and Townsend (1928), and take into account the additional effects of the gas temperature neglected in low current density discharges. The equations are referred to as

(i) the plasma balance equation (cf. eqn. (16)),
(ii) the ion generation equation (cf. eqn. (17)),
(iii) the lateral ion current equation (cf. eqn. (24)),
(iv) the arc current equation, and
(v) the energy balance equation.

From these relations the particle temperature and concentration distributions and the potential distribution about the axis of the arc may be calculated in terms of the coefficients of mobility, drift and ionization. Experimental evidence in support of the Tonks–Langmuir theory is reviewed by Druyvesteyn and Penning (1940).

The above relations have recently been extended by Champion (1952) to the case of the high pressure positive column, and to various intermediate conditions corresponding to low pressures and high electron densities and high pressures with a low degree of ionization. In all cases Champion discusses the influence of a

strong magnetic field; and this is particularly important in the two distinct cases of the vertical and horizontal arc. From the point of view of the energy balance equation the action of the magnetic field is to reduce the losses due to conduction and convection; thus the effect of the magnetic field is not the same in its effect upon the convection losses in a vertical arc as in a horizontal arc.

In addition to the effects produced by the application of an external magnetic field, there are the effects due to the self-magnetic field of the arc current itself: in high-current arcs this self-magnetic field becomes an important discharge parameter. As the current increases, the magnetic field reduces the radial drift of the electrons and the radial electric field, which maintains ambi-polar diffusion at low currents, decreases practically to zero. The axial electron density distribution is then largely determined by the self-magnetic field. This was treated systematically by Tonks (1939), but recently there have been two contributions to the theory. Thonemann and Cowhig (1951) have derived an expression for the electron density distribution determined by the self-magnetic field in a high-current arc. Equilibrium is assumed between the diffusion forces due to the concentration gradient and the induced self-magnetic field. Experimental support is obtained from studies of the mercury arc. Again, Blackman (1951) has considered the theory of the self-magnetic field and compares it with the simpler and more approximate theory of Thonemann and Cowhig. The two theories are in good agreement. Somerville *et al.* (1948) have made a detailed study of the properties of glow discharges in transverse magnetic fields, and have found under certain conditions negative resistance characteristics which may have useful application.

Labouret (1949) has reviewed and classified investigations on the temperature and current density and voltage gradient in the arc column and has treated previous theories. He has developed a new theory for the arc column in molecular gases, based upon the assumptions that the arc temperature is determined by the vibrations of diatomic molecules. These molecules are assumed to undergo an energy cycle leaving the ground state, passing successively through each energy level as a result of molecular impact until they are finally ionized; at this instant radiation is emitted and recombination takes place and energy rapidly degenerates to heat. Froome (1948) has discussed the factors which control the initial rate of increase of current in the arc-like discharges obtained when a condenser of several millifarads is discharged through a small tube containing rarified gas. It is suggested that at higher pressures ionization is probably produced thermally, but that at lower pressures ionization takes place by direct electron collision, the cathode spot having a negligible retarding effect except possibly at very low pressures.

5.1.2. *Cathode effects.*

Recent investigations of electrode phenomena in arc discharges have been concerned with the mechanism of the liberation of electrons which initiate and maintain the arc, and also with the processes of erosion of the electrode material which takes place.

The production of electrode vapour in arc and spark discharges, and the resulting erosion of the electrodes, have been studied considerably during recent years (Llewellyn Jones, 1949, 1949 b, Froome 1948, Finkelburg 1948, Williams, Craggs and Hopwood 1949). A theory of the erosion mechanism given by Llewellyn Jones (1946, 1950) is based on theoretical estimates of the areas of

electrode hot-spots obtained by considering the lateral spread of the electrons and positive ions due to diffusion and their space charge. Areas thus found indicated that very high current densities of the order 10^6 A/cm^2 are possible in the arc phase of typical short sparks. The energy input is then sufficient to raise these small areas to a high temperature, and the heat energy is dissipated by conduction through the metal, by radiation from the glowing hot-spot, and by vaporization of a small amount of electrode metal at the hot-spot. The energy balance at such hot-spots for an equilibrium temperature equal to the boiling point of the electrode material was then considered. Neglecting chemical attack, an equation was found relating the volume of electrode material evaporated per spark (i.e. the spark erosion) to the physical properties of the electrode material. The theory was tested with experimental data on the erosion of sparking plugs.

Considerable experimental work has been carried out since 1946 on the shape and behaviour of cathode hot-spots, leading to estimates of the current density. From a study of the behaviour of the cathode spot on an undisturbed surface of mercury and liquid sodium–potassium alloy by means of a Kerr cell enabling short exposure of the order of 10^{-7} sec to be made, Froome (1946, 1947, 1948, 1949, 1950) found that the current density of emission was about 10^6 A/cm^2, and was greater for the alloy than for mercury. It is suggested, therefore, that the values (Druyvesteyn and Penning 1940) of the emission current density of about 10^4 A/cm^2 usually accepted for normal vacuum arcs is too low. Measurements of the area of the arc cathode spot from measurements of width of tracks left on metals by a cathode spot moving in a magnetic field made by Cobine and Gallagher (1948) again indicate that the current density is greater than values given by previous workers. In arcs passing 2·6 A the following values are given in air at atmospheric pressure with various cathodes: with Cu 124 000 A/cm^2; Al 29 500 A/cm^2; W 74 000 A/cm^2; Hg on Cu 20 700 A/cm^2. The Cu, Al and W electrodes were first oxidized in a bunsen flame. With currents exceeding 5 A the current density could not be measured because of the formation of multiple spots. Somerville and Blevin (1949) and Somerville, Blevin and Fletcher (1952) have also found current densities $\sim 10^6$ A/cm^2 at the cathode spots of transient arcs on aluminium, copper, magnesium, nickel, tin and tungsten cathodes in air at atmospheric pressure. Observations on the shape and size of the cathode hot-spots (Froome 1948, 1949, 1950 with moving arcs, Somerville *et al.* 1952 with stationary arcs) show they are clearly defined and nearly circular. As the current increases, the area at first increases, then splits up into a number of smaller spots. Melted areas on the cathode are greater than the actual area through which the current entered the metal because of thermal conduction. The theoretical estimates of Llewellyn Jones (1946, 1950) are 10^6 A/cm, while Wasserrab (1951) suggests, also on theoretical grounds, than even higher current densities may occur. The realization since 1946, from experimental and theoretical estimates, that current densities attain values of the order of 10^6 A/cm^2, instead of the previously accepted estimates of the order of 4×10^3 A/cm^2 (Guntherschulze 1922), has led to new ideas concerning the mechanism of the cathode electron emission. This is well illustrated by a consideration of the important case of the mercury arc (Wasserrab 1951).

Emission from the liquid mercury cathode was at first considered to be thermionic (Stark 1903), but through the work of Langmuir (1923) and Stolt (1925) was replaced by the Schottky theory (1923) of field-intensified thermionic

emission. The accepted observed current densities at the measured value of the electric field could not be accounted for in terms of the Fowler–Nordheim theory (1928) (see § 4). Slepian (1926) had introduced the idea of the so-called vapour arc, in which the region of electron and ion production occurred in a layer of gas immediately above the cathode. This view was recently discussed by Weizel, Rompe and Schon (1940), but still, however, does not give a completely satisfactory explanation. If the value of the voltage drop V in the cathode dark space is taken as 10, and used in eqn. (45), the field X at the cathode surface given by

$$X = 4V/3d \qquad \ldots\ldots(47)$$

becomes 10^6 to 10^7 v/cm when current densities are 10^6 A/cm^2. These values of X show that with the current densities observed by Froome the field emission of electrons, in accordance with the Fowler–Nordheim theory, now becomes important. These conditions have been treated mathematically by Mackeown (1929). However, the application of the simple Fowler–Nordheim theory is itself open to objection, for it is important to know the exact nature of the surface state of the cathode before the mechanism of emission can be assessed (see § 4). The difficulty of finding a complete explanation is further illustrated by the recent experiments of Weizel *et al.* (1951), who has shown that it is possible to obtain arcs apparently without any observable cathode spots.

5.1.3. *Heavy current arcs.*

Heavy current arcs in hydrogen at low pressures in transverse magnetic fields have been studied by Allen (1951). Pulsed discharges, of 2·5 msec duration, were produced at a known instant in a sinusoidal pulsed magnetic field of 10 msec duration (Champion 1950). An increase in the effective impedance of the arc (ratio of arc p.d. drop to arc current) due to the magnetic field was observed, as would be expected from the reduced mobility of the charged particles arising from the cycloidal motion of charged particles in the crossed electric and magnetic fields. The power dissipated in the discharge to maintain a given current was found to increase with magnetic field strength. Measurements (Allen 1952) of the minimum pressure required for the operation of stable pulsed high-current arcs in hydrogen and air in strong longitudinal magnetic fields have been compared with earlier results (Burhop, Massey and Page 1949) obtained with helium using arcs of 6 A; the minimum pressure was $3{\cdot}2 \times 10^{-3}$ mm Hg of hydrogen in magnetic fields exceeding 1475 gauss, compared with $2{\cdot}05 \times 10^{-4}$ mm Hg of air in magnetic fields exceeding 2000 gauss with an arc potential drop of 190 v (Allen 1952) and $2{\cdot}1 \times 10^{-3}$ to $3{\cdot}1 \times 10^{-3}$ mm Hg of helium in magnetic fields exceeding 1900 gauss with a potential drop of 200 v (Burhop, Massey and Page 1949).

Experimental data on the characteristics of heavy impulse current sparks at the higher pressures were obtained by Higham and Meek (1950 a, b) who investigated long spark channels between point electrodes in various gases. The variation with time of the potential gradient along the spark channel was estimated for discharges in air, nitrogen, oxygen and hydrogen at pressures equal to and lower than atmospheric pressure. The potential gradients were independent of the peak current, except in hydrogen, but were influenced by the rate of decay of the current. On the other hand, the measured areas of cross section of spark channels in air and nitrogen bore a linear relation to the peak currents, and were little influenced by the rate of current decay.

It can be seen from this brief outline that it is still not possible to give a complete theoretical picture of all the fundamental processes of the arc discharge. Considerable work on various aspects such as temperature distribution, electron density distribution, spectral properties and electrode effects is now in progress in various countries. The bibliography is extensive, and only a selected number of references are given here: *Particle Distribution and Radiation Phenomena* : Saufferer (1952), Ecker (1951), Macker and Peters (1951), Edels (1952), Chaghtai (1950), and Steinberg (1951). *Electrode Phenomena* : Thouret and Weizel (1952), Llewellyn Jones (1953 a, b), Finkelnburg (1951).

5.1.4. *Short arcs at electrical contacts.*

The production of an arc by circuit breaking and its consequences have been reviewed by Holm (1946) and more recently by Llewellyn Jones (1949 b, 1953 a), especially for the case of low-power circuits. The explosion or boiling of the microscopic metal bridge, formed in the final stage of the opening operation of the electrodes, provides a ready-made hot-spot from which thermionic emission can take place ; consequently the production of an arc in these circumstances is readily understandable. In any subsequent self-quenching of the arc the state of the cathode surface is of great importance (Llewellyn Jones 1946, 1949 b).

The occurrence of minute arc-like discharges at the close *approach* of contact members operating at low voltages (~50 v) has been the subject of recent work at the Bell Telephone Laboratories, U.S.A. Germer and Haworth (1948, 1949) have shown that such micro-discharges result in the rapid dissipation, mainly at the anode, of the energy stored in the local capacitance, leading to melting and evaporation of a small volume of anode material (see § 5.1.2), and they stress the importance of these hitherto unsuspected effects in the erosion of telephone relay contacts.

Germer and Haworth suggest that a rapid emission of electrons from the cathode is produced by the very close approach of a positive ion to the cathode in the presence of an already high electric field ; the necessary positive ions are assumed to be produced by the ionization of absorbed gas molecules at the anode. A possible theoretical explanation of the high efficiency of electron extraction has been given by Newton (1948). Germer and Haworth (1949), Germer (1951), and Germer and Smith (1952) have found that with certain metals the probability of arcing is increased when the cathode is covered by an insulating film of organic vapour or by insulating particles. They suggest that this can be explained by a charging up by positive ions from the anode of the outer surface of the insulating film on the cathode; the efficiency of the positive ions in drawing out electrons is said then to be greatly increased. Experiments carried out by Haworth (1950) support this view; these experiments were performed on arcs which were struck *in vacuo* between widely spaced electrodes by the deliberate positive ion charging of an insulating film on the cathode. The presence of oxide impurities on the cathode has been shown to lead to a reduction in the arc striking potential in hydrogen (Newman 1951). In an explanation of the mechanism of this phenomenon the experiments on the field emission of electrons from cathodes covered with tarnish layers (§ 4) are relevant (Llewellyn Jones 1953 a, b).

5.2. *Plasma Oscillations*

5.2.1. *Oscillations in glow discharges.*

When a hot cathode discharge is passed through a gas at low pressure, oscillations with frequencies of order of 150 Mc/s or more may be generated. The oscillations are associated with an abnormally large spread of energies in the primary electron beams (Penning 1926, Tonks and Langmuir 1929a). No complete theory of the phenomenon has yet been given, but the measured frequencies are approximately the same as those calculated for the electrostatically controlled vibrations of electrons in the plasma assuming positive ions to be at rest. Neglecting electron pressure, which is usually permissible (Seeliger and Steenbeck 1939), and collisions between electrons and molecules, and denoting electron displacement by s, the equation of motion of a displaced electron is

$$m\,d^2s/dt^2 + 4\pi n_- e^2 s = 0, \qquad \ldots\ldots(48)$$

and the frequency of vibration ν is given by

$$\nu = (n_- e^2/\pi m)^{1/2} = 8980\, n_-^{1/2}\ \text{sec}^{-1}. \qquad \ldots\ldots(49)$$

In these conditions the electron oscillations in a uniform plasma do not give rise to progressive waves. The primary electron beams traversing the oscillating region, can, however, have a disturbance with the plasma electron frequency impressed on them, and carry this to other parts of the tube, where it may be conveyed to an internal or external electrode and through this to a wavemeter. The spread of energy in the primary beams is largely due to the oscillating fields in the plasma, but may be partly due to traversal of regions of the discharge where less regular disturbances are occurring, producing high-frequency noise. Plasma electron oscillations may be recorded without the aid of a primary beam if they occur at a boundary of the plasma.

Merrill and Webb (1939) have made a detailed investigation of a discharge with a flat indirectly heated cathode. By using a probe whose position could be controlled accurately, Merrill and Webb showed that oscillations with large amplitude were only detectable in narrow sheets, near each of which the primary beam acquired an additional range of energies. Some indication was obtained that, for the highest currents, the primary electrons emerged from the cathode sheath with abnormal energies. Merrill and Webb pointed out that primary electrons will be subject to velocity modulation when they pass through an oscillating sheet of plasma and will bunch subsequently, but were unable to obtain quantitative agreement between theory and experiment. Studies of the dependence on discharge conditions of wavelength, and of energy communicable to an external circuit have recently been made by Armstrong (1947) and Neill (1949), Armstrong and Emeléus (1949). Further work has been carried out by Neill and Emeléus (1951), and attention is called to the possibility that turbulent flow (Emeléus 1951) may occur, and that there may be coupling of mechanical and electrical vibrations in the gas (Armstrong, Emeléus and Neill 1951).

A theory of electron oscillations of an unbounded plasma neglecting collisions, and treating in detail only ion gases with a continuous distribution of velocities, which decreases monotonically with increasing velocity, was proposed by Bohm and Gross (1949a). The theory has been extended to include the effects of

collisions and the consideration of more general velocity distributions, showing how these can bring about excitation and damping of plasma oscillations (Bohm and Gross 1949 b).

5.2.2. *Noise.*

Cobine and Gallagher (1947 a) have described work on the generation of noise and oscillations in hot cathode arcs : the random noise produced has a bandwidth of several megacycles per second. The noise is apparently due to positive ions oscillating in two distinct regions of the discharge, viz. the plasma and the potential minimum at the cathode. The disturbance produced by the ion motions appear as voltage variations between the electrodes. Oscillations generated in the plasma are usually below 400 kc/s, while cathode oscillations are about 700 kc/s. As the arc discharge current is increased, the amplitude of the plasma oscillation increases while its frequency decreases. The random noise also depends on the current, increasing rapidly as the Townsend discharge changes into an arc. When the arc is fully established, the noise changes slightly with current. Probe studies show that the noise voltage between probe and cathode increases rapidly as space charge conditions change to permit acceleration of electrons to the probe. An explanation of the noise generation may be found in the random fluctuations of space charge at the electrodes. Calculations based on this theory give voltage fluctuations of the observed order of magnitude. Further studies of noise and oscillation in gas discharges when placed in a magnetic field have been described by Cobine and Gallagher (1947 b).

Measurements of the low-frequency noise spectrum have been made by Martin and Woods (1952) for the case of hot-cathode discharges in mercury vapour saturated at room temperature. The spectrum is continuous, with superposed peaks as previously noted by Cobine and Gallagher. The peaks can be partly attributed to harmonic components of ionic relaxation oscillations. The upper frequency limit of the continuum exceeds 10 Mc/s in some instances. Most of the noise from the tubes studied originated near the cathode. Some relations have been traced between the noise and oscillations and position on the current–voltage characteristic curve for the tube, but there is variation from one tube to another. The generation of high-frequency oscillations by electrical discharges in air at low pressures has been investigated by observations with several types of discharge tubes at frequencies near 200 Mc/s and below 2 Mc/s by Labrum and Bigg (1952). Random noise of high intensity was observed in both these frequency ranges, and was found to be associated with a field reversal in the discharge. Oscillations were also observed at the lower frequencies, and there was evidence that these were the cause of the striations of the positive column. An interesting account of noise generation is found in acoustical testing (Peterson 1939, Dudley, Riesz and Watkins 1939).

5.3. *Spectroscopic Phenomena and Ion Concentrations*

The excitation of atoms in pulsed discharges at atmospheric pressure has been investigated by Craggs, Williams and Hopwood (1948) and Williams, Craggs and Hopwood (1949) by measuring the relative intensities of certain spectral lines. The transport of metal vapour was also investigated in the same experiments. The radial distribution of ion concentrations in spark channels in hydrogen have been measured by observing the Stark broadening of Balmer

lines using Holtsmark's theoretical analysis of the Stark effect (Craggs and Hopwood 1947 a, b). The radial distribution of the light across an arc channel (final stage of a spark discharge) has been examined by Higham and Meek (1950 a, b), and from this estimates of the average ion concentration were made, although the employment of non-uniform fields and impulsive e.m.f's introduced some difficulties of interpretation.

The afterglows observed with long spark channels in hydrogen, oxygen, nitrogen, argon, air and methane have been examined by Craggs, Hopwood and Meek (1947). The light emitted in the hydrogen afterglow was attributed to electron–ion recombination, and the coefficient of recombination deduced appears to be in agreement with the theoretical value. Craggs and Hopwood (1947 b) consider that the recombination of H^+ and electrons to the ground state is sufficient to account for the observed values, but the experiments of Biondi and Brown (1949 b) do not support this view. Edels and Craggs (1951 a) have measured the relative intensities of the first three Balmer series lines excited in low current arcs in hydrogen at pressures from 1 to 2 atmospheres. The populations of the excited states then deduced were not in accordance with a Boltzmann distribution of excited states with a certain mean energy : thus the data could not be described in terms of a definite single excitation temperature for the particular discharge. Preliminary measurements of ion concentrations deduced from Stark broadening of lines emitted by hydrogen arcs indicate that the electron temperature was about 5000°K (Edels and Craggs 1951 b) ; estimates of the gas temperatures were also made.

5.4. *Effects of Irradiation of the Discharge*

When a low-frequency electrodeless discharge of between 50 and 500 c/s is established in a halogen it was found by Joshi and others (Joshi and Varasimham 1940, Joshi and Deshmukh 1941, Joshi and Deo 1943, Deo 1944) that the r.m.s. value of the discharge current decreases on irradiating the vessel with light from an incandescent lamp. This reduction was also observed in oxygen, nitrogen and hydrogen. Harries and von Engel (1951) established that in chlorine the discharge consisted of a large number of pulses per half cycle of the field, and the influence of irradiating different regions of the discharge on the pulse height was investigated. These experiments appeared to be consistent with the view that the chlorine molecules, which form adsorbed layers several molecules thick on the glass walls, are photo-dissociated, and the atoms of chlorine thus formed remove electrons by attachment. The reduction of pulse height owing to the preceding pulse is probably due to dissociation of chlorine molecules by metastable chlorine atoms and resonance radiation.

A comprehensive list of references is given by Arnikar (1952).

REFERENCES

ABDELNABI, I., and MASSEY, H. S. W., 1953, *Proc. Phys. Soc.* A, **66**, 288.
ALLEN, N. L., 1951, *Proc. Phys. Soc.* B, **64**, 276 ; 1952, *Ibid.*, **65**, 697.
ALLIS, W. P., and MORSE, P. M., 1931, *Z. Phys.*, **70**, 567.
ARMSTRONG, E. B., 1947, *Nature, Lond.*, **160**, 713.
ARMSTRONG, E. B., and EMELÉUS, K. G., 1949, *Proc. Instn. Elect. Engrs.*, **96**, Pt. III, 390.
ARMSTRONG, E. B., EMELÉUS, K. G., and NEILL, T. R., 1951, *Proc. R. Irish Acad.* A, **54**, 291.
ARNIKAR, H. J., 1952, *J. Chem. Phys.*, **20**, 917.

BABAT, G., 1947, *J. Instn. Elect. Engrs.*, **94**, Pt. III, 27.
BAILEY, V. A., 1937, *Phil. Mag.*, **23**, 929.
BAILEY, V. A., and MARTYN, D. F., 1934, *Phil. Mag.*, **18**, 369.
BAILEY, V. A., SMITH, R. A., LANDECKER, K., HIGGS, A. J., and HIBBERD, F. H., 1952, *Nature, Lond.*, **169**, 911.
BARTHOLOMCYCZYCK, W., 1940, *Ann. Phys., Lpz.*, **116**, 235.
BATES, D. R., 1950, *Phys. Rev.*, **77**, 718.
BECKEN, O., and SOMMERMEYER, K., 1936, *Z. Phys.*, **102**, 551.
BIONDI, M. A., and BROWN, S. C., 1949 a, *Phys. Rev.*, **75**, 1700 ; 1949 b, *Ibid.*, **76**, 1697.
BLACKMAN, M., 1951, *Proc. Phys. Soc.* B, **64**, 1039.
BOHM, D., and GROSS, E. B., 1949 a, *Phys. Rev.*, **75**, 1851 ; 1949 b, *Ibid.*, **75**, 1864.
BOYD, R. L. F., 1950 a, *Nature, Lond.*, **165**, 142; 1950 b, *Proc. Phys. Soc.* A, **63**, 543.
BRASEFIELD, C. J., 1931, *Phys. Rev.*, **37**, 82.
BROWN, S. C., and MCCARTHY, J. J., 1948, *Rev. Sci. Instrum.*, **19**, 851.
BROWN, S. C., and MACDONALD, A. D., 1949, *Phys. Rev.*, **76**, 1692.
BROWN, S. C., and ROSE, D. J., 1952, *J. Appl. Phys.*, **23**, 711 ; 1952, *Ibid.*, **23**, 719.
BRUCE, C. E. R., 1948, *Nature, Lond.*, **161**, 521.
BRUCE, F. M., 1951, *J. Brit. Instn. Radio Engrs.*, **11**, 121.
BURHOP, E. H. S., MASSEY, H. S. W., and PAGE, G., 1949 (see Guthrie and Wakerling, Chap. 5).
BUTT, A. G., 1933, *World Radio*, April, 28.
CABRERA, N., and MOTT, N. F., 1949, *Rep. Prog. Phys.*, **12**, 163 (London : Physical Society).
CHAGHTAI, M. S. B., 1950, *Proc. Phys. Soc.* B, **63**, 768.
CHAMPION, K. S. W., 1950, *Proc. Phys. Soc.* B, **63**, 795; 1952, *Ibid.*, **65**, 329 and 345.
CHENOT, M., 1948, *Ann. Phys., Paris*, **3**, 277.
COBINE, J. D., and EASTON, E. C., 1943, *J. Appl. Phys.*, **14**, 321.
COBINE, J. D., and GALLAGHER, C. J., 1947 a, *J. Appl. Phys.*, **18**, 110 ; 1947 b, *J. Franklin Inst.*, **243**, 41 ; 1948, *Phys. Rev.*, **74**, 1524.
COBINE, J. D., and WILBUR, D. A., 1951, *J. Appl. Phys.*, **22**, 835.
COMPTON, K. T., 1931, *Phys. Rev.*, **37**, 1077.
COOPER, R., 1947, *J. Instn. Elect. Engrs.*, **94**, Pt. III, 315.
COUSINS, S. W., and WARE, A. A., 1951, *Proc. Phys. Soc.* B, **64**, 159.
CRAGGS, J. D., and HOPWOOD W., 1947 a, *Proc. Phys. Soc.*, **59**, 755 ; 1947 b, *Ibid.*, **59**, 771.
CRAGGS, J. D., HOPWOOD, W., and MEEK, J. M., 1947, *J. Appl. Phys.*, **18**, 919.
CRAGGS, J. D., WILLIAMS, G. C., and HOPWOOD, W., 1948, *Phil. Mag.*, **39**, 329.
CUTOLO, M., 1947, *Nature, Lond.*, **160**, 834 ; 1950, *Ibid.*, **166**, 98.
DEO, P. G., 1944, *Proc. Indian Acad. Sci., Bangalore*, A, **19**, 117.
DRUYVESTEYN, M. J., 1936, *Nature, Lond.*, **137**, 580.
DRUYVESTEYN, M. J., and PENNING, F. M., 1940, *Rev. Mod. Phys.*, **12**, 87.
DUDLEY, H., RIESZ, R. R., and WATKINS, S. S. A., 1939, *J. Franklin Inst.*, **227**, 739.
DUTTON, J., HAYDON, S. C., and LLEWELLYN JONES, F., 1951, *Nature, Lond.*, **168**, 140; 1952 a, *Proc. Roy. Soc.* A, **213**, 203; 1952 b, *Nature, Lond.*, **170**, 601; 1953 a, *Proc. Roy. Soc.* A, in the press; 1953 b, *Brit. J. Appl. Phys.*, in the press.
ECKER, G., 1951, *Z. Phys.*, **130**, 585.
EDELS, H., 1952, *Proc. Phys. Soc.* B, **65**, 794.
EDELS, H., and CRAGGS, J. D., 1951 a, *Proc. Phys. Soc.* A, **64**, 562; 1951 b, *Ibid.*, **64**, 574.
ELENBAAS, W., 1951, *The High Pressure Mercury Vapour Discharge* (Amsterdam : The North Holland Publishing Co.).
EMELÉUS, K. G., 1951, *Proc. Phys. Soc.* B, **64**, 166.
VON ENGEL, A. H., and FRANCIS, G., 1950, *Proc. Phys. Soc.* B, **63**, 823.
FAN, H. Y., 1939, *Phys. Rev.*, **55**, 769.
FELICI, N. J., and MARCHAL, Y., 1948, *Rev. Gen. Elect.*, **57**, 155.
FINKELNBURG, W., 1946, F.I.A.T. Final Report 1052; 1948, *Phys. Rev.* **74**, 222; 1951, *Trans. Amer. Inst. Elect. Engrs.*, **70**, 800.
FISHER, L. H., and BEDERSON, B., 1949, *Phys. Rev.*, **75**, 1324 and 1615; 1951, *Ibid.*, **81**, 109.
FOWLER, R. H., and NORDHEIM, L., 1928, *Proc. Roy. Soc.* A, **119**, 229.
FRANCIS, V. J., and JENKINS, H. G., 1940, *Rep. Prog. Phys.*, **7**, 230 (London : Physical Society).

FROOME, K. D., 1946, *Nature, Lond.*, **157**, 446 ; 1947, *Ibid.*, **159**, 129 ; 1948, *Proc. Phys. Soc.*, **60**, 424 ; 1949, *Ibid.* B, **62**, 805 ; 1950, *Ibid.* B, **63**, 377.
GANGER, H., 1949, *Arch. Elektrotech.*, **39**, 508.
GEBALLE, R., and HARRISON, M. A., 1952, *Phys. Rev.*, **85**, 273.
GERMER, L. H., 1951, *J. Appl. Phys.*, **22**, 955 and 1133.
GERMER, L. H., and HAWORTH, F. E., 1948, *Phys. Rev.*, **73**, 1121; 1949, *J. Appl. Phys.*, **20**, 1085; 1951, *Ibid.*, **22**, 955.
GERMER, L. H., and SMITH, J. L., 1952, *J. Appl. Phys.*, **23**, 553.
GILL, E. W. B., and DONALDSON, R. H., 1931, *Phil. Mag.*, **12**, 719.
GILL, E. W. B., and VON ENGEL, A. H., 1948, *Proc. Roy. Soc.* A, **192**, 446 ; 1949, *Ibid.*, **197**, 112.
VON GUGELBERG, H. L., 1947, *Helv. Phys. Acta*, **20**, 250 and 307.
GUNTHERSCHULZE, A., 1922, *Z. Phys.*, **11**, 74.
GUNTHERSCHULZE, A., and BÄR, R., 1937, *Z. Phys.*, **106**, 668.
GUNTHERSCHULZE, A., and FRICKE, H., 1933, *Z. Phys.*, **86**, 820.
GUTHRIE, A., and WAKERLING, R. K., 1949, *The Characteristics of Electrical Discharges in Magnetic Fields* (New York : McGraw-Hill).
GUTTON, C., 1924, *C. R. Acad. Sci., Paris*, **178**, 467.
GUTTON, C., and GUTTON, H., 1928, *C. R. Acad. Sci., Paris*, **186**, 303.
GUTTON, C., MITRA, S. K., and YLOSTALO, V., 1923, *C. R. Acad. Sci., Paris*, **176**, 1871.
HALE, D. H., 1948, *Phys. Rev.*, **73**, 1046.
HARRIES, W. L., and VON ENGEL, A. H., 1951, *J. Chem. Phys.*, **19**, 514; 1951, *Proc. Phys. Soc.* B, **64**, 915.
HATCH, A. J., and WILLIAMS, H. B., 1951, Conference on Gaseous Electronics (New York : American Physical Society).
HAWORTH, F. E., 1950, *Phys. Rev.*, **80**, 223 ; 1951, *J. Appl. Phys.*, **22**, 606.
HAYMAN, R. L., 1929, *Phil. Mag.*, **7**, 586.
HEIDEMANN, E., 1931, *Phys. Rev.*, **37**, 978.
HERLIN, M. A., and BROWN, S. C., 1948 a, *Phys. Rev.*, **74**, 291 ; 1948 b, *Ibid.*, **74**, 1650.
HIGHAM, J. B., and MEEK, J. M., 1950 a, *Proc. Phys. Soc.* B, **63**, 633 ; 1950 b, *Ibid.*, **63**, 649.
HITTORF, W., 1884, *Wied. Ann.*, **52**, 473.
HOCHBERG, B., and SANDBERG E., 1942, *J. Tech. Phys., U.S.S.R.*, **12**, 65.
HOLM, R., 1924, *Phys. Z.*, **25**, 497; 1946, *Electrical Contacts* (Stockholm : Almquist and Wiksells).
HOLSTEIN, T., 1946 a, *Phys. Rev.*, **69**, 50 (A) ; 1946 b, *Ibid.*, **70**, 367.
HOLT, R. B., RICHARDSON, J. M., HOWLAND, B., and MCCLURE, B. T., 1950, *Phys. Rev.*, **77**, 239.
HORNBECK, J. A., 1951, *Phys. Rev.*, **84**, 615.
HORTON, F., and MILLEST, D. M., 1946, *Proc. Roy. Soc.* A, **185**, 381.
HOWELL, A. H., 1939, *Trans. Amer. Inst. Elect. Engrs.*, **58**, 193.
HUXLEY, L. G. H., 1950, *Proc. Roy. Soc.* A, **200**, 486.
HUXLEY, L. G. H., FOSTER, H. G., and NEWTON, C. C., 1948, *Proc. Phys. Soc.*, **61**, 134.
HUXLEY, L. G. H., and ZAAZOU, A. A., 1949, *Proc. Roy. Soc.* A, **196**, 402.
JENKINS, R. O., 1943, *Rep. Prog. Phys.*, **9**, 177 (London : Physical Society).
JOHNSON, P., 1930, *Phil. Mag.*, **10**, 921.
JOHNSON, R. A., MCCLURE, B. T., and HOLT, R. B., 1950, *Phys. Rev.*, **80**, 376.
JOSHI, S. S., and DEO, P. G., 1943, *Nature, Lond.*, **151**, 561.
JOSHI, S. S., and DESHMUKH, G. S., 1941, *Nature, Lond.*, **147**, 806.
JOSHI, S. S., and VARASIMHAN, V., 1940, *Curr. Sci.*, **9**, 536.
JURRIAANSE, T., and DRUYVESTEYN, M. J., 1935, *Physica*, **3**, 825.
KACHICKAS, G. A., and FISHER, L. H., 1952, *Phys. Rev.*, **88**, 878.
KENTY, C., 1928, *Phys. Rev.*, **32**, 624.
KEYSTON, J. E., 1933, *Phil. Mag.*, **16**, 625.
KIRCHNER, F., 1925, *Ann. Phys., Lpz.*, **77**, 187.
KNIPP, C. T., and KNIPP, J. K., 1931, *Phys. Rev.*, **38**, 948.
KRASIK, S., ALPERT, D., and MCCOUBREY, A. O., 1949, *Phys. Rev.*, **76**, 722.
LABOURET, J., 1949, *Rev. Gen. Elect.*, **58**, 357.
LABRUM, N. R., and BIGG, E. K., 1952, *Proc. Phys. Soc.* B, **65**, 356.
LANGMUIR, I., 1923, *Science*, **58**, 290; 1923, *Rev. Gen. Elect.*, **26**, 731; 1928, *Z. Phys.*, **46**, 283.
LASSEN, H., 1931, *Arch. Elektrotech.*, **25**, 322.

LLEWELLYN JONES, F., 1931, *Phil. Mag.*, **11**, 163; 1933, *Ibid.*, **15**, 958; 1936, *Proc. Phys. Soc.*, **48**, 513; 1939, *Phil. Mag.*, **28**, 192; 1944, *Proc. Phys. Soc.*, **56**, 239; 1946, *Nature, Lond.*, **157**, 371 and 480; 1949 a, *Proc. Phys. Soc.* B, **62**, 366; 1949 b, *Proc. Instn. Elect. Engrs.*, **96**, 305; 1950, *Brit. J. Appl. Phys.*, **1**, 60; 1951, *Nature, Lond.*, **168**, 140; 1952, *Ibid.*, **169**, 960; 1953 a, *The Fundamental Processes of Electrical Contact Phenomena* (H.M.S.O. : London : Special Report No. 24); 1953 b, *Proc. Instn. Elect. Engrs.*, Pt. I, in the press.

LLEWELLYN JONES, F., and DAVIES, D. E., 1951 a, *Proc. Phys. Soc.* B, **64**, 397; 1951 b, *Ibid.*, **64**, 519.

LLEWELLYN JONES, F., and MORGAN, C. G., 1951, *Phys. Rev.*, **82**, 970; 1953, *Proc. Roy. Soc.* A, in the press.

LLEWELLYN JONES, F., and MORGAN, G. D., 1951 a, *Proc. Phys. Soc.* B, **64**, 560 ; 1951 b, *Ibid.*, **64**, 574.

LLEWELLYN JONES, F., and PARKER, A. B., 1950, *Nature, Lond.*, **165**, 960 ; 1952, *Proc. Roy. Soc.* A, **213**, 185.

LLEWELLYN JONES, F., and DE LA PERRELLE, E. T., 1951, *Nature, Lond.*, **168**, 160; 1953, *Proc. Roy. Soc.* A, **216**, 267.

LLEWELLYN JONES, F., DE LA PERRELLE, E. T., and MORGAN, C. G., 1951, *C. R. Acad. Sci., Paris*, **232**, 592.

LLEWELLYN JONES, F., and WILLIAMS, G. C., 1953 a, *Proc. Phys. Soc.* B, **66**, 17; 1953 b, *Ibid.*, **66**, 345.

LOEB, L. B., 1939, *Fundamental Processes of Electrical Discharges in Gases* (New York : Wiley), Chap. IX, p. 407, ref. 62; 1951, *Phys. Rev.*, **81**, 287.

LOEB, L. B., and MEEK, J. M., 1941, *The Mechanism of the Electric Spark* (Stanford : University Press), pp. 27–30.

LOEB, L. B., and WIJSMAN, R. A., 1948, *J. Appl. Phys.*, **19**, 797.

LUNT, R. W., VON ENGEL A. H., and MEEK, J. M., 1941, *Rep. Prog. Phys.* **8**, 338 (London : Physical Society).

MACCALLUM, S. P., KLATZOW, L., and KEYSTON, J. E., 1933, *Phil. Mag.*, **16**, 193.

MACKEOWN, S. S., 1929, *Phys. Rev.*, **34**, 611.

MACKER, H., and PETERS, T., 1951, *Z. Phys. Chem.*, **198**, 318.

MACKINNON, K. A., 1929, *Phil. Mag.*, **8**, 605.

MARGENAU, H., 1946, *Phys. Rev.*, **69**, 508 ; 1948 a, *Ibid.*, **73**, 326 ; 1948 b, *Ibid.*, **73**, 297.

MARTIN, H., and WOODS, H. A., 1952, *Proc. Phys. Soc.* B, **65**, 281.

MASCH, K., 1932, *Arch. Elektrotech.*, **26**, 589.

MASSEY, H. S. W., 1949, *Rep. Prog. Phys.*, **12**, 248 (London : Physical Society); 1950, *Negative Ions* (Cambridge : University Press); 1952, *Advances in Physics* (*Phil. Mag.* Suppl.), **1**, 395.

MASSEY, H. S. W., and BURHOP, E. H. S., 1952, *Electronic and Ionic Impact Phenomena* (Oxford : Clarendon Press).

MASSEY, H. S. W., and MOHR, C. B. O., 1934, *Proc. Roy. Soc.* A, **144**, 188.

MEDICUS, G., 1948, *Z. angew. Phys.*, **1**, 106.

MEEK, J. M., 1940, *Phys. Rev.*, **57**, 722; 1942, *J. Instn. Elect. Engrs.*, **89**, 335.

MEEK, J. M., and PERRY, F. R., 1946, *Rep. Prog. Phys.*, **10**, 314 (London : Physical Society).

MERRILL, H. J., and WEBB, H. W., 1939, *Phys. Rev.*, **55**, 1191.

MEYEROTT, R., 1944, *Phys. Rev.*, **66**, 242.

MOAK, C. D., REESE, H., and GOOD, W. M., 1951, *Nucleonics*, **9**, 18.

MORGAN, G. D., 1953, *Sci. Progr.*, **161**, 22.

MUNSON, R. J., and TYNDALL, A. M., 1941, *Proc. Roy. Soc.* A, **177**, 187.

NEILL, T. R., 1949, *Nature, Lond.*, **163**, 59.

NEILL, T. R., and EMELÉUS, K. G., 1951, *Proc. R. Irish Acad.* A, **53**, 197.

NEUERT, H., 1949, *Z. Naturforsch.*, **4a**, 449.

NEWMAN, R., 1951, *Brit. J. Appl. Phys.*, **2**, 303.

NEWTON, R. R., 1948, *Phys. Rev.*, **73**, 1122.

NIELSEN, R. A., 1936, *Phys. Rev.*, **50**, 950.

NISEWANGER, C. R., HOLMES, J. R., and WEISSLER, G. L., 1946, *J. Opt. Soc. Amer.*, **36**, 581.

OLIPHANT, M. L., and MOON, P. B., 1930, *Proc. Roy. Soc.* A, **127**, 388.

PAAVOLA, M., 1929, *Arch. Elektrotech.*, **22**, 443.

PAETOW, H., 1939, *Z. Phys.*, **111**, 770.

PENNING, F. M., 1926, *Physica*, **6**, 241 ; 1928, *Proc. K. Acad. Wet. Amst.*, **31**, 14 ; 1928, *Physica*, **8**, 13 ; 1931, *Phil. Mag.*, **11**, 961 ; 1937, *Physica*, **4**, 71.
PETERSON, E., 1939, *Bell Lab. Record*, **18**, 81.
PHELPS, A. V., and BROWN, S. C., 1952, *Phys. Rev.*, **86**, 102.
PHELPS, A. V., FUNDINGSLAND, O. T., and BROWN, S. C., 1951, *Phys. Rev.*, **84**, 559.
POSIN, D. Q., 1948, *Phys. Rev.*, **73**, 496.
PROWSE, W. A., 1950, *Proc. Brit. Instn. Radio Engrs.*, **10**, 333.
PROWSE, W. A., and JASINSKI, W., 1951, *Proc. Instn. Elect. Engrs.*, **98**, Pt. IV, 101.
RAMBERG, W., 1932, *Ann. Phys., Lpz.*, **12**, 314.
RATCLIFFE, J. A., and SHAW, I. J., 1948, *Proc. Roy. Soc.* A, **193**, 311.
REUKEMA, L. E., 1927, *Trans. Amer. Inst. Elect. Engrs.*, **46**, 1314 ; 1928, *Ibid.*, **47**, 38.
ROHDE, L., 1932, *Ann. Phys., Lpz.*, **12**, 569.
RUTHERGLEN, J. S., and COLE, J. F. I., 1947, *Nature, Lond.*, **160**, 545.
SANDERS, F. G., 1932, *Phys. Rev.*, **44**, 667.
SAUFFERER, H., 1952, *Z. Phys.*, **131**, 376.
SAYERS, J., 1938, *Proc. Roy. Soc.* A, **169**, 83 ; 1943, *Rep. Prog. Phys.*, **9**, 52 (London: Physical Society) ; 1947, *Nature, Lond.*, **159**, 117 ; 1950, *Proc. Conf. Ionos. Phys.*, State College, Pennsylvania.
SCHADE, R., 1937, *Z. Phys.*, **104**, 487.
SCHNITGER, H., 1935, *Z. Phys.*, **96**, 551 ; 1936, *Ibid.*, **102**, 163.
SCHOTTKY, W., 1923, *Z. Phys.*, **14**, 63 ; 1924, *Phys. Z.*, **25**, 342.
SEELIGER, R., and BECKEN, O., 1935, *Ann. Phys., Lpz.*, **96**, 551.
SEELIGER, R., and STEENBECK, M., 1939, *Z. Phys.*, **113**, 526.
SHAW, I. J., 1951, *Proc. Phys. Soc.* B, **64**, 1.
SKILLING, H. H., 1939, *Trans. Amer. Inst. Elect. Engrs.*, **58**, 161.
SKILLING, H. H., and BRENNER, W. C., 1941, *Trans. Amer. Inst. Elect. Engrs.*, **60**, 112.
SLATER, J. G., 1946, *Rev. Mod. Phys.*, **18**, 481.
SLEPIAN, J., 1926, *Phys. Rev.*, **27**, 407.
SMITH, C. G., 1941, *Phys. Rev.*, **59**, 997 ; 1947, *Ibid.*, **71**, 135.
SOMERVILLE, J. M., 1952, *Proc. Phys. Soc.* B, **65**, 620.
SOMERVILLE, J. M., and BLEVIN, W. R., 1949, *Phys. Rev.*, **46**, 982.
SOMERVILLE, J. M., BLEVIN, W. R., and FLETCHER, N. H., 1952, *Proc. Phys. Soc.* B, **65**, 963.
SOMERVILLE, J. M., CHAMPION, K. S. W., and BIGG, E. K., 1948, *Aust. J. Sci. Res.*, **1**, 400.
STARK, J., 1903, *Ann. Phys., Lpz.*, **12**, 673.
STEINBERG, R. K., 1951, *J. Appl. Phys.*, **21**, 1028.
STERN, T. E., GOSSLING, B. S., and FOWLER, R. H., 1928, *Proc. Roy. Soc.* A, **124**, 699.
STOLT, H., 1925, *Z. Phys.*, **31**, 240.
STUECKELBERG, E. C. G., and MORSE, P. M., 1930, *Phys. Rev.*, **36**, 16.
TELLEGEN, B. D. H., 1933, *Nature, Lond.*, **131**, 840.
THOMSON, J., 1930, *Phil. Mag.*, **10**, 280 ; 1934, *Ibid.*, **18**, 696 ; 1937, *Ibid.*, **23**, 1.
THOMSON, J. J., 1927, *Phil. Mag.*, **4**, 1128 ; 1928, *Proc. Phys. Soc.*, **40**, 79.
THONEMANN, P. C., and COWHIG, W. T., 1951, *Proc. Phys. Soc.* B, **64**, 345.
THONEMANN, P. C., MOFFATT, J., ROAF, D., and SANDERS, J. H., 1948, *Proc. Phys. Soc.*, **61**, 483.
THOURET, W., and WEIZEL, W., 1952, *Z. Phys.*, **131**, 170.
TILLES, A., 1934, *Phys. Rev.*, **46**, 1015.
TONKS, L., 1934, *Phys. Rev.*, **46**, 278 ; 1935, *Ibid.*, **48**, 562 ; 1939, *Ibid.*, **56**, 360.
TONKS, L., and LANGMUIR, I., 1929 a, *Phys. Rev.*, **33**, 195 ; 1929 b, *Ibid.*, **34**, 876.
TOWNSEND, J. S., 1915, *Electricity in Gases* (Oxford : Clarendon Press) ; 1923, *Phil. Mag.*, **45**, 444 ; 1924, *Motions of Electrons in Gases* (Oxford : Clarendon Press) ; 1928, *C. R. Acad. Sci., Paris*, **186**, 55 ; 1930, *Phil. Mag.*, **9**, 1145 ; 1931, *Ibid.*, **11**, 1112 ; 1932, *Ibid.*, **13**, 745 ; 1933, *Ibid.*, **16**, 729.
TOWNSEND, J. S., and DONALDSON, R. H., 1928, *Phil. Mag.*, **5**, 178.
TOWNSEND, J. S., and LLEWELLYN JONES, F., 1931 a, *Phil. Mag.*, **11**, 679 ; 1931 b, *Ibid.*, **12**, 815.
TOWNSEND, J. S., and NETHERCOT, W., 1929, *Phil. Mag.*, **7**, 600.
TRUMP, J. G., CLOUD, R. W., MANN, J. G., and HANSON, G. P., 1950, *Trans. Amer. Inst. Elect. Engrs.*, **69**, 961.

TRUMP, J. G., SAFFORD, F. J., and CLOUD, R. W., 1941, *Trans. Amer. Inst. Elect. Engrs.*, **60**, 132.
TYKOCINSKI-TYKOCINER, S., 1932, *Phil. Mag.*, **13**, 953.
TYNDALL, A. M., and POWELL, C. F., 1931, *Proc. Roy. Soc.* A, **134**, 125.
VALLE, G., 1950, *Nuovo Cim.*, **7**, 174; 1952, *Ibid.*, **9**, 145.
VARNERIN, L. J., 1951, *Phys. Rev.*, **84**, 563.
WARE, A. A., 1951, *Phil. Trans. Roy. Soc.* A, **243**, 197.
WASSERAB, T., 1951, *Z. Phys.*, **130**, 311.
WEHRLI, M., 1927, *Z. Phys.*, **44**, 301.
WEIZEL, W., ROMPE, R., and SCHON, M., 1940, *Phys. Z.*, **115**, 179.
WEIZEL, W., and THOURET, W., 1952, *Z. Phys.*, **131**, 170.
WERNER, S., 1950, *Nature, Lond.*, **165**, 1018.
WILLIAMS, G. C., CRAGGS, J. D., and HOPWOOD, W., 1949, *Proc. Phys. Soc.* B, **63**, 633.
WOLF, K., 1939, *Z. Phys.*, **112**, 96 and 148.
YARNOLD, G. D., 1932, *Phil. Mag.*, **13**, 1179.
YOUNG, B. R., 1950, *J. Appl. Phys.*, **21**, 222.
ZEIER, O., 1932, *Ann. Phys., Lpz.*, **14**, 415.
ZELENY, J., 1942, *J. Appl. Phys.*, **13**, 444.
ZUBER, K., 1925, *Ann. Phys., Lpz.*, **76**, 231.

FLUCTUATION THEORY IN PHYSICAL MEASUREMENTS

By C. W. McCOMBIE
Department of Natural Philosophy, University of Aberdeen

CONTENTS

PAGE

§ 1. Introduction 267
§ 2. Measurements with suspended systems 267
2.1. Measurements involving a single reading 268
2.1.1. Uncertainty in couple 268
2.1.2. Uncertainty in quantity producing the couple .. 269
2.2. Measurements involving a number of readings 271
2.2.1. Correlation function 271
2.2.2. Measurements involving two observations 275
2.2.3. Procedures involving averaging 277
2.3. Feedback and amplification 280
§ 3. Random forces : correlation and power spectra 281
3.1. Outline of random force method 281
3.2. Power spectrum 282
3.3. Experimental determination of power spectrum of a random force 284
3.4. Probability distributions 285
3.5. Fluctuations in circuit with inductance and resistance 285
3.5.1. Power spectrum method 285
3.5.2. Correlation function method 286
3.6. Time integrals of current and voltage 287
3.7. Measurement of a current with an ideal amplifier 288
3.8. Fluctuations of a torsionless suspension 289
3.9. Further mathematical considerations 289
§ 4. Physical basis 291
4.1. Equilibrium systems : equipartition 292
4.2. Equilibrium systems : random forces 294
4.3. Nyquist's result 296
4.4. Extension of Nyquist's result to an impedance 297
4.5. Recent discussion of Nyquist's results 298
4.6. Upper limit to frequencies for which Nyquist's formula is valid 299
4.7. Systems acted upon by a number of independent fluctuating forces 300
4.8. Systems not in equilibrium 301
§ 5. Feedback 304
5.1. Feedback with suspended systems 304
5.2. General considerations on feedback 305
§ 6. Radiation detectors 306
6.1. Fluctuations in energy transfer 306
6.2. The thermocouple 307
6.3. Thermocouple as a radiation detector : chopping of radiation 308
6.4. The bolometer 310
6.5. Limitations on radiation detectors in general 311
§ 7. Discrimination against noise 314
7.1. Wiener's criterion 314
7.2. Optimum response of critically damped galvanometer 316
7.3. Optimum response to a pulse 317
7.4. Information theory 318
References 319

ABSTRACT. An attempt is made to give a coherent, elementary account of the ways in which fluctuation theory has been applied to some of the simpler types of physical measurement. Uncertainties in measurements involving suspended systems are discussed on the basis of simple correlation function arguments, and consideration is given to the methods of measurement appropriate when there are various practical limitations on the parameters of the suspended system. There follows an elementary development of the mathematical and physical considerations usually employed in treating fluctuations in linear measuring instruments: remarks on the status of the random force method as applied to equilibrium systems are included. The circumstances under which there is an absolute limit to the attainable accuracy are discussed. It is shown that, in the case of measurements with a suspended system, feedback may enable these limits to be attained with a convenient measurement procedure. An account is given of various approaches to the calculation of the limits of accuracy in measurements with radiation detectors. Finally some elementary results are established concerning the optimum characteristics of an instrument used to follow a varying signal in the presence of noise.

§1. INTRODUCTION

THE existence of spontaneous fluctuations sets an absolute limit to the accuracy which can be obtained when measuring a physical quantity in a prescribed way. This paper reviews the methods available for predicting the mean square errors introduced by fluctuations into a variety of measurements, and discusses the choice of measurement procedures that will make these errors as small as is possible under given conditions. Our treatment is intended to be elementary, and complete in itself. Considerable space has been devoted to the underlying mathematical methods and physical principles: a number of the points discussed in this connection have been the subject of lively comment in recent literature.

To preserve the elementary nature of the discussion and keep the length of the review within reasonable bounds we have considered only linear measuring systems. We have also omitted explicit discussion of the details of fluctuations in electronic amplifiers (noise): elementary treatments are readily available in textbooks, and surveys of more advanced aspects are presented in works in the Massachusetts Institute of Technology Radiation Laboratory Series, to which reference will be made later. Some other topics not treated here are discussed in the reviews by MacDonald (1948) and van der Ziel (1952).

The formal development of the basis of the subject is preceded by a discussion of the effect of fluctuations on measurements with suspended systems. This makes the considerations applicable to this most fundamental type of measurement as accessible as possible; it also provides us with a simple and convincing demonstration of the importance of the correlation function, a concept which figures largely in the subsequent sections.

Our discussion of a measurement procedure aims always at determining the root mean square error in the estimate of the quantity being measured. It did not seem necessary to regard the somewhat vague concept of 'minimum detectable signal' as fundamental; it is clear that the root mean square error in the estimate gives the order of magnitude of the minimum signal that can be detected with the given procedure.

§2. MEASUREMENTS WITH SUSPENDED SYSTEMS

Many physical measurements consist essentially in observing the deflection produced by a couple acting on a suspended system. It is well known that such a system will fluctuate about its equilibrium position, the fluctuations

becoming large enough to be observed if the restoring force is made sufficiently weak. The mean square magnitude of the deflection from the position of mechanical equilibrium can at once be predicted by regarding the whole as a system with one degree of freedom and applying the equipartition principle: the justification for this procedure will be discussed later (§ 4.1). On considering the mean potential energy of the suspended system, we get

$$\tfrac{1}{2}c\overline{\Delta\theta^2}=\tfrac{1}{2}kT, \qquad \ldots\ldots(1)$$

where c is the torsional constant of the suspension, $\Delta\theta$ the fluctuation in the deflection, k Boltzmann's constant, and T the absolute temperature. This result is taken to hold independently of the amount of damping present. It has been checked quantitatively by Kappler (1931).

We propose to discuss in this section the effect of such fluctuations on the reproducibility of estimates of the magnitude of the couple, or of the quantity of which the couple is a measure, e.g. the current in the case of the galvanometer.

2.1. *Measurements involving a Single Reading*

2.1.1. *Uncertainty in couple.*

Let us suppose to begin with that we know the torsional constant c and have determined the zero position by previous averaging, and that we wish to determine the magnitude of a couple applied to the system by taking a single reading when the system has had time to reach equilibrium with the couple applied. The mean square error in the estimate of the couple so obtained is easily found, if it is assumed that the couple introduces no new fluctuations: the fluctuations of the system predicted by (1) are then simply superimposed on the response to the applied couple. If θ, the observed deflection, is measured from the zero point, the estimate of the couple is $c\theta$; this will, of course, be in error because of the fluctuation $\Delta\theta$ included in θ. The error ΔP in the estimate arising from a fluctuation $\Delta\theta$ in the deflection is given by $\Delta P=c\Delta\theta$, so that $\overline{\Delta P^2}=c^2\overline{\Delta\theta^2}$ (the average being taken over a large number of repetitions of the measurement) and, by (1),

$$\overline{\Delta P^2}=ckT. \qquad \ldots\ldots(2)$$

The stiffer the suspension the bigger the uncertainty in the couple. As we shall see later, however, this is necessarily so only if we confine ourselves to a single reading.

The time required to make the measurement is usually important and in what follows *we shall always assume that the measurement must be completed in a prescribed time* S. Reducing c may then involve changing the moment of inertia I or the damping constant κ in order that the response of the suspended system may be sufficiently rapid. We are usually limited in practice by difficulty in effecting these auxiliary changes rather than by difficulty in reducing c. Suppose, for example, that the measurement procedure consists in critically damping the suspension, i.e. making

$$\kappa=2(Ic)^{1/2}, \qquad \ldots\ldots(3)$$

and taking a reading of the deflection θ from the zero position after a time equal to the undamped period τ_0 given by

$$\tau_0=2\pi(I/c)^{1/2}, \qquad \ldots\ldots(4)$$

which is taken to be the response time. Clearly τ_0 must be made equal to S. The estimate of the couple will then be $c\theta/0{\cdot}986$ and the mean square error $\overline{\Delta P^2}$ in the estimate will be $c^2\Delta\theta^2/(0{\cdot}986)^2$, or $c^2\Delta\theta^2$ to a sufficiently good approximation. If now c is reduced, I must be reduced to keep τ_0 equal to S according to (4), and κ must be reduced to keep the damping critical according to (3). We therefore require all the parameters c, I and κ to be small : depending on which of these proves to be the limiting factor, various expressions for $\overline{\Delta P^2}$ will be of interest. Using (3) and (4) we obtain

$$\overline{\Delta P^2} = ckT = \pi\kappa kT/S = 4\pi^2 IkT/S^2. \qquad \ldots\ldots(5)$$

The different forms of dependence of $\overline{\Delta P^2}$ on S should be noted.

The limits of accuracy given in (5) refer only to the particular measurement procedure discussed. It is convenient to introduce here, for purposes of comparison, a limit to the accuracy of a measurement with a suspended system which cannot be surpassed by any measurement procedure. We shall later (§3.8) present reasons for believing that if the damping constant is κ the suspended system must be regarded as being acted upon by a rapidly fluctuating couple which is such that if we take its time average over each of a large number of intervals of length S, the mean square value of these time averages will be $2\kappa kT/S$. Now if we require to make in a time S a measurement of an applied couple acting on the system, the most complete experimental data we could possibly acquire from observations on the suspended system would be the value of the total couple (applied couple plus fluctuating couple) at each instant in the interval. The best estimate of the applied couple would then be given by the average of this total couple over the time S. The error ΔP in the estimate would be just the average of the fluctuating couple over the time S and so $\overline{\Delta P^2}$ would be given by

$$\overline{\Delta P^2} = 2\kappa kT/S. \qquad \ldots\ldots(6)$$

Clearly no other measurement procedure occupying a time S can give a smaller $\overline{\Delta P^2}$, and measurement procedures are to be judged by how closely they allow us to make $\overline{\Delta P^2}$ approach $2\kappa_{\min}kT/S$, where $\kappa_{\min}$ is the minimum attainable value for the damping. The shift of emphasis from c to κ should be noted.

We see that with the critical damping procedure discussed $\overline{\Delta P^2}$ exceeds the optimum by a factor of only $\pi/2$ provided κ is the limiting factor. According to (3) and (4), the present method always requires that

$$I = \kappa S/4\pi \qquad \ldots\ldots(7)$$

and the method can be used with κ equal to $\kappa_{\min}$ (i.e. κ is the limiting factor) only if $I_{\min}$ does not exceed $\kappa_{\min}S/4\pi$. If it does we must increase κ to make the damping critical and $\overline{\Delta P^2}$ will in the absence of limitations on c be further increased by a factor equal to the ratio of $I_{\min}$ to $\kappa_{\min}S/4\pi$. One of the aims of the discussion of more complicated measurement procedure to be taken up in §2.2 will be to find methods of approaching the optimum limit when $I_{\min}$ is much greater than $\kappa_{\min}S$, i.e. methods which enable us to take advantage of arbitrarily small values of $\kappa_{\min}$, in spite of restrictions on I.

2.1.2. *Uncertainty in quantity producing the couple.*

Usually, of course, one is not interested in the couple as such, but rather in some force, current, or the like, which is producing it, and to which it is proportional. One may then be able to increase the ultimate sensitivity of the

system without decreasing c, simply by increasing the constant of proportionality. Very often, however, the situation is complicated by the existence of a relationship between the constant of proportionality and the constants of the suspended system. If, for example, we are primarily interested in measuring a force, the constant of proportionality is essentially the arm of the couple, denoted by a, and it may in practice be impossible to increase this without increasing κ and I. If the force is denoted by F, $\Delta P = a\Delta F$ and we get for the critical damping procedure discussed

$$\overline{\Delta F^2} = (c/a^2)\boldsymbol{k}T = (\kappa/a^2)\pi\boldsymbol{k}T/S = (I/a^2)4\pi^2\boldsymbol{k}T/S^2, \qquad \ldots\ldots(8)$$

so that whichever of c/a^2, κ/a^2, and I/a^2 proves to be the limiting factor should be made as small as possible. It may not be desirable to have c small.

An interesting and important example is afforded by the galvanometer. To be definite we shall discuss the moving-coil type, but similar considerations apply to the moving-magnet instrument. We assume that the fluctuations satisfy (1) even when a steady current is maintained in the coil by some non-equilibrium device, such as a battery. This assumption may not hold if the circuit contains non-metallic resistors (see §4.8). The couple acting on the coil is related to the current i by the equation $P = Gi$, where G is the effective flux linkage of the coil, which can, by changing the strength of the magnetic field, be varied (in practice up to some limiting value) without altering the suspended system. Increasing the constant of proportionality G, however, increases the damping, for if R is the circuit resistance and κ_0 the air damping the total damping is given by $\kappa = \kappa_0 + G^2/R$. The overall effect when I and c are readjusted to give critical damping is easily found. Combining the second expression for ΔP^2 in (5) with $\Delta P = G\Delta i$, and using the expression just given for κ, we get

$$\overline{\Delta i^2} = \left(\frac{\kappa_0}{G^2} + \frac{1}{R}\right)\frac{\pi}{S}\boldsymbol{k}T. \qquad \ldots\ldots(9)$$

Increasing G diminishes the first term in the bracket (i.e. makes the air damping relatively less important) and so makes $\overline{\Delta i^2}$ smaller. However, with the measurement procedure under discussion and fixed circuit resistance R there is, according to (9), a minimum value for $\overline{\Delta i^2}$, viz.

$$\overline{\Delta i^2} = \pi\boldsymbol{k}T/RS. \qquad \ldots\ldots(10)$$

Comparing this with the second expression for $\overline{\Delta P^2}$ in (5) we see that prescribing $1/R$ in this case is analogous to prescribing κ in the case of measuring a couple. The analogy is even closer than might appear, for just as there is a fluctuating couple associated with damping κ, there is a fluctuating voltage associated with a resistance R (see §§3.5 and 3.6) ; the average of this over a time S has mean square value $2R\boldsymbol{k}T/S$. Consequently an estimate of a voltage in a circuit of resistance R made in a time S will have a mean square error of at least $2R\boldsymbol{k}T/S$, and the corresponding minimum error in the estimate of a current will be $2\boldsymbol{k}T/RS$. The limit indicated by (10) is only $\pi/2$ times as large as this. Since air damping is usually negligible for a galvanometer with critical damping, this limit can be attained, provided only that R does not exceed the maximum critical damping resistance which is determined by the minimum value of I/G^2 since by (7) we require for critical damping $I = (G^2/R)S/4\pi$: if it does exceed this value we must either shunt the galvanometer, thus reducing the circuit resistance and

so increasing the uncertainty in current, or introduce air damping, which again increases the uncertainty. This difficulty is essentially the same as that which may arise from limitations on I when a couple is being measured, and we shall consider in § 2.2 measurement procedures which in spite of limitations on I/G^2 enable us to approach the optimum limit when the circuit resistance is arbitrarily large.

The optimum value of $\overline{\Delta i^2}$ given by (10) varies inversely as the response time: high accuracy can be obtained only at the expense of taking a long time over the measurements. To indicate a typical order of magnitude we may suppose that the circuit resistance is 100 ohm and the observation time one second. Then according to (10) the root mean square error in the estimate of the current will be $1{\cdot}1 \times 10^{-11}$ amp at room temperature.

The very important result (10) was obtained by Ising (1926). If one knows the d.c. sensitivity of the critically damped galvanometer as well as R and the undamped period, one can use (10) to obtain $\overline{\Delta\theta^2}$. In this way Ising was able to show that the fluctuations in the zero point observed by Moll and Burger (1925) in the primary galvanometer of their thermo-relay amplifier could be largely accounted for without postulating external disturbances.

2.2. *Measurements involving a Number of Readings*

Very often the zero position of the suspension cannot be assumed to remain unchanged and has to be redetermined each time a measurement is made. The simplest measurement procedure will then involve two observations of the deflection from an arbitrary zero, one to determine the undeflected and one to determine the deflected position. The uncertainty in the estimate depends on the mean square fluctuation in the difference of these two deflections. The calculation of this fluctuation is complicated by the fact that the fluctuations in the two deflections will not in general be independent. It is, for example, obvious from the general nature of records of observed fluctuations (fig. 1) that the mean square fluctuation in the difference of the observations can be made very small by making the time interval between successive observations small; the fluctuation in deflection then tends to be the same at the two instants of observation. A more striking example is afforded by a very lightly damped suspension. As can be seen from fig. 1, the fluctuation then takes the form of a sinusoidal oscillation slowly varying in amplitude and phase; the period is the period of the suspension. The fluctuation deflection tends to be the same at two instants separated by a period, and to have exactly opposite values at instants half a period apart.

To be able to deal with these complications which arise in all measurements involving more than a single reading we must follow Zernike (1932) and make use of the correlation function.

2.2.1. *Correlation function.*

The correlation appears in our calculations in a simple way. Let us suppose that deflections θ_1 and θ_2 measured from an arbitrary zero are observed at times t_1 and t_2. Then the mean square fluctuation in $\theta_2 - \theta_1$ is given by

$$\overline{(\Delta\theta_2 - \Delta\theta_1)^2} = \overline{\Delta\theta_1^2} + \overline{\Delta\theta_2^2} - 2\overline{\Delta\theta_1 \Delta\theta_2}. \qquad \ldots\ldots(11)$$

Here $\overline{\Delta\theta_1^2}$ and $\overline{\Delta\theta_2^2}$ are equal, and are just the $\overline{\Delta\theta^2}$ given by (1). If $\Delta\theta_1$ and $\Delta\theta_2$ were independent, the last term would be zero and the mean square fluctuation

in the difference would then be just twice the mean square fluctuation in a single reading. However, this does not hold if the two fluctuation deflections are not independent. If the deflections tend to be the same, $\Delta\theta_1\Delta\theta_2$ will be positive and will approach $\overline{\Delta\theta^2}$ and, from (11), $\overline{(\Delta\theta_2-\Delta\theta_1)^2}$ will approach zero; $\overline{\Delta\theta_1\Delta\theta_2}$ cannot exceed $\overline{\Delta\theta^2}$ since $(\Delta\theta_2-\Delta\theta_1)^2$ must be positive. Similarly, if the deflections tend to have opposite sign $\overline{\Delta\theta_1\Delta\theta_2}$ will be negative but cannot be less than $-\overline{\Delta\theta^2}$. In the limit when $\overline{\Delta\theta_1\Delta\theta_2}$ equals $-\overline{\Delta\theta^2}$ the mean square error in the difference will be four times the mean square error in a single observation.

The value of $\overline{\Delta\theta_1\Delta\theta_2}$ for a given form of suspension will clearly depend only on the modulus of the difference between the two observation times. Denoting this difference by τ, we define the correlation function $\psi(\tau)$ by

$$\psi(\tau)=\overline{\Delta\theta(t_1)\Delta\theta(t_1+\tau)}. \qquad \ldots\ldots(12)$$

It will be seen that this function gives all the information about the form of the fluctuations necessary for a discussion of problems in which we are interested. For example, (11) can at once be rewritten

$$\overline{(\Delta\theta_2-\Delta\theta_1)^2}=2\{\psi(0)-\psi(t_2-t_1)\}. \qquad \ldots\ldots(13)$$

It is useful to define also a normalized correlation function $r(\tau)$ by

$$r(\tau)=\psi(\tau)/\overline{\Delta\theta^2}, \qquad \ldots\ldots(14)$$

in which case $\overline{(\Delta\theta_2-\Delta\theta_1)^2}$ may be written $2\overline{\Delta\theta^2}\{1-r(t_2-t_1)\}$. The very general discussion given so far leads to the following properties of $\psi(\tau)$:

$$\psi(0)=\overline{\Delta\theta^2}\,..(a), \quad -\overline{\Delta\theta^2}\leqslant\psi(\tau)\leqslant+\overline{\Delta\theta^2}\,..(b), \quad \psi(\tau)=\psi(-\tau)\,..(c). \qquad \ldots\ldots(15)$$

We require, however, an expression for $\psi(\tau)$ in terms of the constants of the suspension; the basis of its derivation is twofold.

In the first place, regarding the suspension as a system of one degree of freedom, we get from statistical mechanics the equipartition result (1) already used, and also the fact that the fluctuation in angular velocity at a given time is independent of the fluctuation in deflection at that time. This last result follows from the fact that the probability of the fluctuation in the deflection lying between $\Delta\theta$ and $\Delta\theta+d(\Delta\theta)$ and the fluctuation in the angular velocity lying between $\Delta\omega$ and $\Delta\omega+d(\Delta\omega)$ is proportional to

$$[\exp\{(-\tfrac{1}{2}c\Delta\theta^2-\tfrac{1}{2}I\Delta\omega^2)/kT\}]\;d(\Delta\theta)d(\Delta\omega).$$

Clearly, if we consider only instants for which $\Delta\theta$ has some definite value, the distribution of values of $\Delta\omega$ will be proportional to $\{\exp(-\tfrac{1}{2}I\Delta\omega^2/kT)\}\,d(\Delta\omega)$, quite independently of the value of $\Delta\theta$.

In the second place we require an assumption about the form of the mechanisms causing the fluctuations. We assume that their action on the suspension takes the form of a very rapidly fluctuating couple, of mean value zero, which acts in addition to the ordinary damping force. The assumption of rapid fluctuation enables us to assume that subsequent to any given instant the history of the random couple is quite independent of the fluctuation in the deflection at that instant: this would not be so, for example, if the fluctuating couple varied so slowly that the fluctuation in the deflection was always approximately proportional to the fluctuating couple.

To proceed with the derivation, we now suppose that the fluctuations in deflection and angular velocity at time t_1 are $\Delta\theta_1$ and $\Delta\omega_1$ respectively. Since the system is linear we can regard the fluctuation deflection at a subsequent

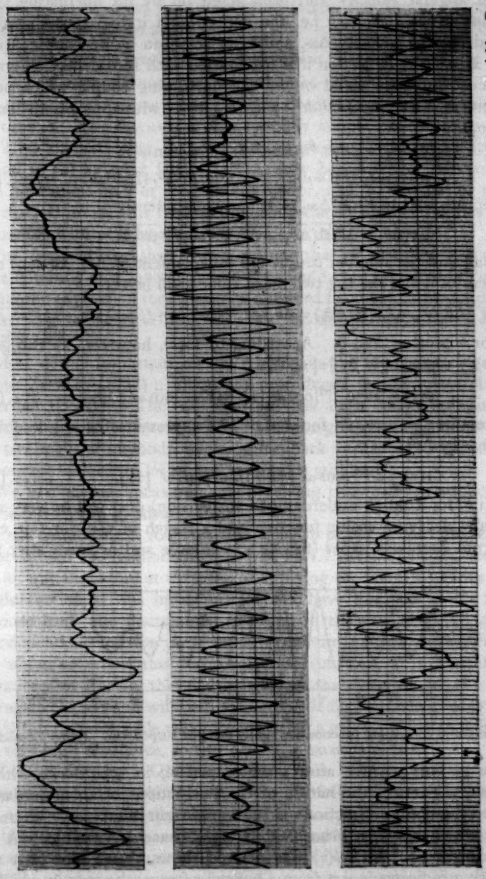


Fig. 1. Records of fluctuations of a galvanometer (free period 2·2 sec) with varying amounts of electromagnetic damping. (*After* R. V. Jones.) The galvanometer is overdamped (short-circuited) in (*a*), underdamped (open-circuited) in (*b*), and critically damped in (*c*). The fluctuations of the primary light spot are amplified by photocells used in conjunction with a critically damped secondary galvanometer (period 1·6 sec).

instant t_2 as being made up of three parts : (i) the residue of $\Delta\theta_1$, $\mathfrak{R}(\Delta\theta_1)$ i.e. the deflection the suspension would have had if it had started at time t_1 with deflection $\Delta\theta_1$ and zero angular velocity and been acted on by no fluctuating couple; (ii) the residue of $\Delta\omega_1$, $\mathfrak{R}(\Delta\omega_1)$ i.e. the deflection if it had started with zero deflection and angular velocity $\Delta\omega_1$ and been acted on by no couple; (iii) the random part $\mathfrak{R}$, i.e. the deflection if it had started with zero deflection and zero angular velocity but been acted on by the fluctuating couple which actually acts on the system. Solving $Id^2\theta/dt^2+\kappa\, d\theta/dt+c\theta=0$ with appropriate boundary conditions we find

$$\mathfrak{R}(\Delta\theta_1)=\Delta\theta_1 e^{-\beta\tau}(\cos\omega'\tau+(\beta/\omega')\sin\omega'\tau)$$

where $$\tau=t_2-t_1,\quad \beta=\kappa/2I,\quad \omega'=(c/I-\kappa^2/4I^2)^{1/2}. \qquad \ldots\ldots(16)$$

We may write $\Delta\theta_2=\mathfrak{R}(\Delta\theta_1)+\mathfrak{R}(\Delta\omega_1)+\mathfrak{R}$ and consequently

$$\overline{\Delta\theta_1\Delta\theta_2}=\overline{\Delta\theta_1\mathfrak{R}(\Delta\theta_1)}+\overline{\Delta\theta_1\mathfrak{R}(\Delta\omega_1)}+\overline{\Delta\theta_1\mathfrak{R}}.$$

Since both $\Delta\omega_1$ and the random couple are independent of $\Delta\theta_1$, $\mathfrak{R}(\Delta\omega_1)$ and $\mathfrak{R}$ are independent of it and the last two mean values will be zero.

Consequently $\overline{\Delta\theta_1\Delta\theta_2}=\Delta\theta_1\mathfrak{R}(\Delta\theta_1)=\Delta\theta_1^2 e^{-\beta\tau}\{\cos\omega'\tau+(\beta/\omega')\sin\omega'\tau\}$.

This holds for τ greater than 0. According to 15(c), however, we can get the general result by replacing τ by $|\tau|$ so that

$$\psi(\tau)=\overline{\Delta\theta^2}\,e^{-\beta|\tau|}\{\cos\omega'|\tau|+(\beta/\omega')\sin\omega'|\tau|\}. \qquad \ldots\ldots(17)$$

This is the required expression for $\psi(\tau)$. The expression for $r(\tau)$ is obtained on dividing this by $\overline{\Delta\theta^2}$:

$$r(\tau)=e^{-\beta|\tau|}\{\cos\omega'|\tau|+(\beta/\omega')\sin\omega'|\tau|\}. \qquad \ldots\ldots(18)$$

For small β (i.e. for a much underdamped suspension) $\psi(\tau)$ will be very nearly an exponentially decaying cosine function and a sketch of $r(\tau)$ for this case is reproduced from Zernike's paper (fig. 2). It is large and negative after half a

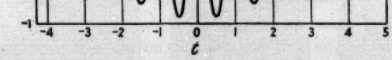


Fig. 2. Normalized correlation function for lightly damped suspension. (*After* Zernicke)

period, and large and positive after a whole period, in accordance with the discussion already given of the nature of the fluctuations of an underdamped system. The correlation function is so important that any alternative interpretation which helps to bring out its significance is valuable. A consideration of the way in which $\psi(\tau)$ was derived shows that a plot of it or of $r(\tau)$ for positive τ indicates, apart from a constant factor, the way in which the suspension returns to its equilibrium position, if it is started from rest in a deflected position and acted on by no fluctuating force. This result will prove useful in applications. Again, if we consider a large number of identical suspensions all of which at a given instant happen to have deflection $\Delta\theta_0$ from their equilibrium position (the angular velocities being assumed to have their equilibrium distribution and the random forces to be acting) the average of the deflections

after a time τ will be just the residue of $\Delta\theta_0$ at that time, the average of the other parts of the deflection being zero, and this residue is just $\Delta\theta_0 r(\tau)$; in other words $r(\tau)$ shows, apart from a constant factor, the way in which the mean of the deflection decays to zero.

2.2.2. *Measurements involving two observations.*

We can now discuss the uncertainties which arise with a number of measurement procedures when they are used to measure a couple or a current in a time S. The discussion of a specially simple procedure given in §§2.1.1 and 2.1.2 suggests that for each procedure considered special attention should be paid to (*a*) by what factor $\overline{\Delta P^2}$ exceeds $2\kappa_{\min}kT/S$ or $\overline{\Delta i^2}$ exceeds $2kT/RS$ when limitations on I or I/G^2 do not matter, and (*b*) whether or not limitations on I or I/G^2 prevent us from approaching the small uncertainties which should be possible with very small values of $\kappa_{\min}$ or $1/R$.

For simplicity we shall always assume that air damping is a negligible part of the damping of a galvanometer. It should then be noted in connection with (*a*) that the expression for $\overline{\Delta i^2}$ can be obtained from the corresponding expression for $\overline{\Delta P^2}$ by replacing κ by $1/R$, since for the galvanometer $\overline{\Delta P^2} = G^2\overline{\Delta i^2}$ and $\kappa = G^2/R$.

Most of the important results relating to the fundamental question (*a*) are due to Zernike (1932) : these are reproduced in the review by Barnes and Silverman (1934). The question (*b*) has been less emphasized but is of considerable practical interest. The discussion given here will be completed in §5.1 by a treatment of methods involving feedback : these are of great interest when a favourable answer to (*b*) as well as to (*a*) is required.

We first discuss measurements involving two readings applicable when the zero is not known in advance.

Critical damping : Response time equal to period of observation. If the zero is not known the method discussed in §2.1.1 can still be applied provided we take an initial reading at the time of applying the couple. Since the residue of the fluctuation at this initial instant will have dropped nearly to zero in the response time, the fluctuations included in the two observed deflections will be independent and the mean square error in the estimate will be just double that obtained when the zero is known (cf. eqn. (5)). In the absence of limitations on I and c this will be π times the optimum for the given $\kappa_{\min}$ but here again the method will be unfavourable if $I_{\min}$ is much greater than $\kappa_{\min}S/4\pi$. Nor will it be favourable when we are using a galvanometer to measure a current in a very high resistance, so that $(I/G^2)_{\min}$ exceeds $S/4\pi R$.

Critical damping : Optimum response time. Suppose now that we drop the requirement that τ_0 should equal S. It follows from our interpretation of $r(\tau)$ as giving essentially the way in which an initial deflection from the equilibrium decays to zero, that if the couple applied at t_1 and $t_2 - t_1$ is S the estimate of the couple will be $c(\theta_2-\theta_1)/[1-r(S)]$. In the present case $r(S)$ is $e^{-x}(1+x)$ where x is written for $2\pi S/\tau_0$: this is obtained from (18) by equating β to $2\pi/\tau_0$ and letting ω' go to zero. Remembering that $\overline{(\Delta\theta_2-\Delta\theta_1)^2}$ is $2\overline{\Delta\theta^2}\{1-r(S)\}$ we find

$$\overline{\Delta P^2} = \frac{c^2\overline{(\Delta\theta_2-\Delta\theta_1)^2}}{\{1-r(S)\}^2} = \frac{\kappa kT}{S}\,\frac{x}{1-e^{-x}(1+x)} = \frac{2IkT}{S^2}\,\frac{x^2}{1-e^{-x}(1+x)}. \quad \ldots\ldots(19)$$

It follows from the second expression that for fixed κ and S, $\overline{\Delta P^2}$ attains its minimum value when $x = 1{\cdot}8$, i.e. when $\tau_0 = 3{\cdot}5S$. Thus we arrive at Zernike's

result that if a critically damped suspension with fixed κ, or a critically damped galvanometer with fixed R, is being used to measure a couple or a current the response time should, if possible (i.e. if the other constants can be given the required values), be adjusted to be about 3·5 times the observation time S and the uncertainties will then be

$$\overline{\Delta P^2} = 3{\cdot}3\kappa \boldsymbol{k} T/S, \qquad \overline{\Delta i^2} = 3{\cdot}3\boldsymbol{k}T/RS. \qquad \ldots\ldots(20)$$

(The interval between the observations will have to be timed exactly, an experimental complication.) If this method can be used with $\kappa = \kappa_{\min}$, $\overline{\Delta P^2}$ is only 3·3/2 times the optimum, a value almost as small as that obtained with the procedure described in §2.1.2 for a suspension of known zero. This accuracy can be attained only if $I_{\min}$ is not greater than $3{\cdot}5\kappa_{\min}S/4\pi$, a value 3·5 times as great as the value for $I_{\min}$ required when we make τ_0 equal to S. If $I_{\min}$ exceeds this value we must, to keep the damping critical, either keep τ_0 equal to 3·5S and increase κ, or keep κ equal to $\kappa_{\min}$ and increase τ_0. It can be shown from (19) that if I and S are fixed, $\overline{\Delta P^2}$ always decreases with decreasing x, i.e. with decreasing κ since $x = \kappa S/2I$, and the second alternative is therefore preferable. If $I_{\min}$ is much larger than $\kappa_{\min}S$ (i.e. x is very small) we find

$$\overline{\Delta P^2} = 4I\boldsymbol{k}T/S^2 \qquad \ldots\ldots(21)$$

and this is about twenty times smaller than the mean square error which would be obtained for the same suspension if the response time were made equal to S. Nevertheless the method is still not favourable for very large $I_{\min}$. Similarly if a current has to be measured in a circuit of very high resistance an improvement in accuracy can be obtained if the requirement $\tau_0 = S$ is dropped, but the critical damping method is still not favourable.

Lightly damped suspension : Period equal to observation time. As another possibility we may, following Zernike (1932), consider using a much underdamped system of period S in the hope of utilizing the tendency of the fluctuations in the deflection of such a system to be the same at instants separated by a period. As to the actual measurement procedure, we must note that in this case it is not advantageous to apply the couple immediately we have made the first observation. The maximum throw is attained *half* a period after applying the couple; therefore in order to get as large a difference between the readings as possible, while preserving the advantage of making the observations a period apart, one should wait half a period after taking the first reading before applying the couple (or current). For very small damping the throw will be to a sufficiently good approximation twice the steady state deflection and the estimate of the couple will be $\frac{1}{2}c(\theta_2 - \theta_1)$. Also $r(S) = \exp(-\beta S)$ and so $\overline{(\Delta\theta_2 - \Delta\theta_1)^2} = 2\overline{\Delta\theta^2}\{1 - \exp(-\beta S)\}$. Hence, using $\kappa \ll 2(Ic)^{1/2}$, $S \simeq 2\pi(I/c)^{1/2}$ and $\beta = \kappa/2I$, we obtain to a good approximation

$$\overline{\Delta P^2} = \pi^2\kappa\boldsymbol{k}T/S \qquad \ldots\ldots(22)$$

and with a galvanometer used in the same way

$$\overline{\Delta i^2} = \pi^2\boldsymbol{k}T/RS. \qquad \ldots\ldots(23)$$

If κ is equal to $\kappa_{\min}$ this value for $\overline{\Delta P^2}$ is about five times the optimum. This mean square error compares unfavourably with the mean square errors which could, in the absence of limitations on I, be obtained with the methods previously described. It is important to note, however, that with the present method the

requirement is for I to be *large*, in fact large enough for the system to be much underdamped when the damping is $\kappa_{\min}$. The limits set to the accuracy of the previously described procedures by a minimum value for I have, therefore, no counterpart in the present procedure. Similarly for a galvanometer used in this way, the value of $\overline{\Delta i^2}$ given by (23) can, in spite of the existence of a minimum value of I/G^2, be attained when the circuit resistance is arbitrarily large. Thus, although the present method is not very favourable in the light of considerations (*a*) of the first paragraph of this section and is not very convenient in practice, it is favourable in the light of consideration (*b*).

Torsionless suspension : *Fluxmeter device.* An important case is that in which we use a suspension with zero torsion constant, in which case the galvanometer becomes a fluxmeter. When the torsion constant is zero, there is no longer any fixed position from which to measure the fluctuations, but $\Delta\theta_2-\Delta\theta_1$ is still significant, being the fluctuation from the initial position in time τ. In this case, substituting (18) in $(\boldsymbol{k}T/c)\{1-r(\tau)\}$ and letting c tend to zero, one obtains

$$\overline{(\Delta\theta_2-\Delta\theta_1)^2}=\frac{2\boldsymbol{k}T}{\kappa}\left[\tau-\frac{I}{\kappa}\left\{1-\exp\left(-\frac{\kappa\tau}{I}\right)\right\}\right]. \qquad \ldots\ldots(24)$$

A more direct derivation of this will be indicated in §3.8. For $\tau\gg I/\kappa$ the expression for $\overline{(\Delta\theta_2-\Delta\theta_1)^2}$ becomes $(2\boldsymbol{k}T/\kappa)\tau$ and for $\tau\ll I/\kappa$ it is $(\boldsymbol{k}T/I)\tau^2$ (i.e. $\overline{(d\theta/dt)^2}\tau^2$ as was to be expected).

If we apply the couple for a time S and observe the initial and final deflections θ_1 and θ_2, the estimate of the couple will be

$$P=\kappa\left[S-\frac{I}{\kappa}\left\{1-\exp\left(-\frac{\kappa S}{I}\right)\right\}\right]^{-1}(\theta_2-\theta_1) \qquad \ldots\ldots(25)$$

and accordingly the mean square error will be given by

$$\overline{\Delta P^2}=2\kappa\boldsymbol{k}T\left[S-\frac{I}{\kappa}\left\{1-\exp\left(-\frac{\kappa S}{I}\right)\right\}\right]^{-1}. \qquad \ldots\ldots(26)$$

The corresponding expression for $\overline{\Delta i^2}$ is obtained in the usual way. If I is much less than κS (or I/G^2 much less than S/R) we find

$$\overline{\Delta P^2}=2\kappa\boldsymbol{k}T/S, \qquad \overline{\Delta i^2}=2\boldsymbol{k}T/RS, \qquad \ldots\ldots(27)$$

and, provided these conditions can be satisfied with $\kappa=\kappa_{\min}$ and R the circuit resistance without shunting, *the optimum accuracy in measuring couple or current can be attained.* On the other hand, where $\kappa_{\min}S$ or S/R is so small that $I_{\min}$ or $(I/G^2)_{\min}$ is relatively large in comparison, the method is less favourable. If $I_{\min}$ is very large we get $\overline{\Delta P^2}=4I\boldsymbol{k}T/S^2$. This agrees with (21) as was to be expected, since that limit also refers to a case in which, I being fixed, κ and c tend to zero.

2.2.3. *Procedures involving averaging.*

Averaging by integration. So far we have discussed cases in which the response time of the suspension is as long as or longer than the observation time, and have confined ourselves to two observations. It is frequently much more practicable to use an instrument of much shorter response time and to average the deflection in some suitable way. This will also ensure that the instrument is ready for further use immediately after the observation period. Suppose, for example, that one knows the zero position of a suspension and wishes to measure a couple

in a time S. We may make the instrument respond so fast (by making the period short and $\kappa/2I$ large enough to prevent protracted oscillations) that the time necessary to reach the steady state may be negligible compared with S. Our best estimate of the couple will be

$$P = \frac{c}{S}\int_0^S \theta(t)\,dt \qquad \ldots\ldots(28)$$

and consequently

$$\overline{\Delta P^2} = \overline{\left\{\frac{c}{S}\int_0^S \Delta\theta(t)\,dt\right\}^2} = \frac{c^2}{S^2}\int_0^S\int_0^S \overline{\Delta\theta(t_1)\Delta\theta(t_2)}\,dt_1\,dt_2. \qquad \ldots\ldots(29)$$

The transformation to a double integral expresses the result in terms of the correlation function. Using the expression for the correlation function (17) and the fact that the range of time over which correlation is appreciable is small compared with S, we get

$$\overline{\Delta P^2} = 2\kappa kT/S. \qquad \ldots\ldots(30)$$

Provided the procedure can be used with κ equal to $\kappa_{\min}$ the optimum limit is attained. This is possible however only if $I_{\min} \ll \kappa_{\min} S$, as the time over which correlation is appreciable (which is also the time taken to attain the final deflection) will not be less than I/κ and consequently this must be made much less than S. Similarly the method enables us to attain the optimum limit with a galvanometer only if, without shunting, S/R is large compared with $(I/G^2)_{\min}$. If I is fixed and κ has to be adjusted to make $I/\kappa \ll S$, it follows on combining this equation with (30) that $\overline{\Delta P^2} \gg 2IkT/S^2$: the method is therefore not favourable if I is the limiting factor (compare, for example, eqn. (21)).

If the zero is not known, one will average over half the time to determine the zero, and then apply the couple and average over the other half to determine the deflected position. The mean square fluctuation in each average is double that of an average over the whole period and, the two averages being independent because of the negligible correlation time, the error in the difference will be twice the error in either average. Thus the mean square fluctuation will be four times that given by the method appropriate when the zero is known. Thus we find, for example,

$$\overline{\Delta i^2} = 8kT/RS. \qquad \ldots\ldots(31)$$

This is less favourable than the critically damped galvanometer method, but has the advantage that the galvanometer is immediately available for further use. It should be noted that if it is possible to apply the current in opposite directions over the two halves of the observation time the optimum result can be attained without knowing the zero.

In practice, of course, one does not observe the integrals but takes a large number of equally spaced readings. Zernike shows that there is little to be gained by making the interval between observations much less than half the free period of the suspension. He also discusses averages over times of the order of magnitude of a free period, and shows that for light damping the averaging should be done over an integral multiple of the free period. This is reasonable when one considers the quasi-periodic nature of the fluctuations in this case. These results can be obtained by correlation function calculations similar to those which we use to deal with other problems, and they need not be discussed further here.

Couple applied for alternate periods. Zernike discusses a modification of the method described in § 2.2.2 for using a lightly damped suspension : it gives the same $\overline{\Delta P^2}$ and has the advantage of leaving the suspension ready for further use. The period is made equal to $S/2$ and readings θ_1, θ_2 and θ_3 are taken at times 0, $S/2$ and S, the couple being applied at $S/4$ and removed at $3S/4$. For small damping the estimate of P is $(-\theta_1+2\theta_2-\theta_3)\ c/4$ and the calculation of $\overline{\Delta P^2}$ is straightforward. This suggests that if a long observation time is available it may be convenient to make the period very short compared with S and to apply the couple for alternate periods, taking a reading at the middle of each period and averaging successive throws from one side to the other. It can be shown, by methods similar to those used below to discuss the resonance procedure, that if the number of periods is large

$$\overline{\Delta P^2}=\gamma\pi^2\kappa\boldsymbol{k}T/S \qquad (1\leqslant\gamma\leqslant 2) \qquad \ldots\ldots(32)$$

where γ is near 2 or 1 according as the correlation over a time S is large or small. Limitations on I can at most increase γ to 2. The suspension will again be ready for further use in a negligible time. It is moreover possible in this method to make a correction for drift.

Couple applied for alternate half periods: Resonance. Finally, a method discussed by Hardy (1930), Firestone (1932) and Van Lear (1933), as well as by Zernike, is of some interest. It consists essentially in using an underdamped (and so resonant) suspension and applying the couple to be measured over alternate half periods. It is easy to show that, if the damping is small, the swing (peak to peak) of the oscillation which eventually builds up will be approximately $4/\beta\tau_0$ times the steady deflection produced by the couple when it acts all the time. The time taken for the oscillation to reach 99% of its final amplitude is approximately $4{\cdot}6/\beta$. If, as is assumed by the authors mentioned, one uses the whole observation period to build up to the full oscillation, so that $S=4{\cdot}6/\beta$, and then takes the difference between the last deflection to the right and the last to the left, i.e. $\theta_L-\theta_R$, one's estimate of the couple is $c(\theta_L-\theta_R)\beta\tau_0/4$. Now θ_L and θ_R are separated by half a period, so that

$$\overline{(\Delta\theta_L-\Delta\theta_R)^2}=2\overline{\Delta\theta^2}\{1+\exp(-\pi\beta/\omega')\}\simeq 4\overline{\Delta\theta^2},$$

for small damping. It follows easily, with the aid of the above expression for S, that

$$\overline{\Delta P^2}=\frac{c^2 4\overline{\Delta\theta^2}\beta^2\tau_0^2}{16}=\frac{23\kappa\boldsymbol{k}T}{S}, \qquad \ldots\ldots(33)$$

and, for the galvanometer with negligible air damping used in this way,

$$\overline{\Delta i^2}=23\boldsymbol{k}T/RS. \qquad \ldots\ldots(34)$$

There is no advantage in adjusting G to make the electromagnetic damping equal the air damping, a condition assumed by Hardy, Firestone and Van Lear. This value for $\overline{\Delta i^2}$ is large, but the present method, like the previous one, does not require small I, and it, too, has the advantage that the presence of drifts will have relatively little effect, as the two observations are made very close together. Zernike shows that one can reduce $\overline{\Delta i^2}$ for a given S by a factor of nearly two by arranging that the amplitude attained is only 78% of the maximum amplitude.

It seems, however, to be somewhat more advantageous to use a system which attains its final amplitude in a time very much less than S (i.e. for which $4{\cdot}6/\beta\ll S$) and then to average the throws from right to left, $(\theta_{L,r}-\theta_{R,r})$, $r=1, 2\ldots N$,

where $(2N-1)\tau_0'/2 = S$, τ_0' being the damped period. The estimate of the couple will then be

$$P = (\tfrac{1}{4}\beta\tau_0 c/N) \sum_{r=1}^{N} (\theta_{L,r} - \theta_{R,r}). \qquad \ldots\ldots(35)$$

Using $r(\tau_0'/2) = -\exp(-\beta\tau_0'/2)$, $r(\tau_0') = \exp(-\beta\tau_0')$, etc., we obtain

$$\overline{\left[\sum_{r=1}^{N} (\Delta\theta_{L,r} - \Delta\theta_{R,r})\right]^2} = \overline{\Delta\theta^2}[2N - 2(2N-1)r(\tau_0'/2) + \ldots\ldots - 2r\{(2N-1)\tau_0'/2\}]$$

$$= 4N\overline{\Delta\theta^2}\left[\frac{1}{1-x} - \frac{1}{2} - \frac{x(1-x^{2N})}{2N(1-x)^2}\right], \; x = \exp(-\beta\tau_0'/2). \qquad \ldots\ldots(36)$$

Since small damping implies $\beta\tau_0' \ll 1$, and $4{\cdot}6/\beta \ll S$ implies $N\beta\tau_0' \gg 1$, this reduces to $4N\overline{\Delta\theta^2}(2/\beta\tau_0')$. Hence, noting that to a sufficiently good approximation $N\tau_0' = S$ and $\tau_0' = \tau_0$, we obtain

$$\overline{\Delta P^2} = \pi^2\kappa\boldsymbol{k}T/S. \qquad \ldots\ldots(37)$$

This gives for the galvanometer

$$\overline{\Delta i^2} = \pi^2\boldsymbol{k}T/RS. \qquad \ldots\ldots(38)$$

These uncertainties are the same as would be obtained with the same system if the couple or current were applied for alternate *periods*, since the fact that correlation over the time S is small implies $\gamma = 1$ in (32). It should be remarked that if it is possible to reverse the couple or current instead of just switching it off, the estimate of the couple in terms of the observed throws will be just half that given in (35) and consequently

$$\overline{\Delta P^2} = \tfrac{1}{4}\pi^2\kappa\boldsymbol{k}T/S, \qquad \overline{\Delta i^2} = \tfrac{1}{4}\pi^2\boldsymbol{k}T/RS. \qquad \ldots\ldots(39)$$

These exceed the optimum values by only 25%. Similar results hold for the alternate period method.

The resonance method as just described leaves the instrument ready for further use, but shares the disadvantage of the integration method in that it is applicable only when $I_{\min} \ll \kappa_{\min} S$ (i.e. when $4{\cdot}6/\beta \ll S$). It is of interest in connection with radiation detectors (cf. §6.3).

2.3. *Feedback and Amplification*

We see that if, with the methods discussed, we wish to measure a couple with a suspension for which $\kappa_{\min} S$ is much less than $I_{\min}$, we can approach the optimum value for $\overline{\Delta P^2}$ only by using an underdamped suspension and that even then $\overline{\Delta P^2}$ will exceed the optimum by a factor of about five (unless it is possible to reverse the couple). Similar considerations apply to a galvanometer when S/R is much less than $(I/G^2)_{\min}$. We shall see in §5.1 that under these circumstances the use of feedback may provide a very convenient way of attaining the optimum limit.

Methods of amplifying the deflection of a suspended system are of great importance: we have seen that even from the point of view of minimizing the errors due to fluctuations, it may not be either necessary or desirable to use a weak suspension. Amplification of the deflection by using a light spot and thermal detectors or photocells has been discussed by many workers, including Moll and Burger (1925), Hill (1934, 1938, 1948, 1949), Jones (1951), Moss (1935), Preston (1946) and Frankenhaeuser and MacDonald (1949). Instruments

suitable for use with the resonance method are described by Hardy (1930) and Firestone (1932). In so far as the response time of the amplifying mechanism is much shorter than that of the suspension, the theory given in this chapter still applies. Discussions of the case in which the response times are comparable are given by Van Lear (1933), Passoth (1942) and Jones and McCombie (1952).

Various authors have described galvanometer amplifiers exhibiting fluctuations not much greater than those to be expected on the assumption that external disturbances are negligible (Moll and Burger 1925, Barnes and Matossi 1932, and Hill 1949). Ising (1932) observed the fluctuations of a galvanometer directly with a microscope and found the r.m.s. deflection to be within 5% of his theoretical value. The galvanometer amplifier built and tested by R. V. Jones (see Jones 1951, and Jones and McCombie 1952) exhibited fluctuations having a root mean square value agreeing to within 1% with the predicted value. The independence of the root mean square deflection of the primary galvanometer on the amount of electromagnetic damping was established (see fig. 1). The instrument is quite capable of being used at a sensitivity at which this limit is important, as its drift over, say, a thousand times the response time was much less than the root mean square fluctuation deflection.

For a discussion of the use of the galvanometer with an electrometer triode reference should be made to the paper by van der Ziel (1942).

§3. RANDOM FORCES: CORRELATION AND POWER SPECTRA

3.1. *Outline of Random Force Method*

The discussion of the fluctuations of a suspended system given in the previous section was based on statistical mechanical arguments supplemented by the assumption that the fluctuations could be regarded as being caused by a fluctuating couple. A qualitative assumption as to the nature of this couple (that it varied rapidly) was sufficient as quantitative results were obtained from statistical mechanics.

Statistical mechanical arguments are, however, not always applicable: thus, for example, an electronic amplifier is clearly not an equilibrium system, nor is there any obvious plausible way in which we can regard it as equivalent to an equilibrium system (as when we ignored the steady current in treating galvanometers). Under these circumstances we resort to what has come to be the standard method of treating fluctuations in mechanical and, more especially, electrical systems. This depends on establishing once and for all rules which enable us, on inspecting such a system, to associate with various parts of it fluctuating forces, couples, voltages or the like with definite statistical properties. These statistical properties have to be sufficient to enable us to derive from them alone the statistical properties of the fluctuations of the system; an example is the rule which associates a fluctuating voltage with each resistance in an electrical circuit, the statistical properties of the voltage depending only on the magnitude of the resistance and its temperature. The next section will be concerned with the physical basis of this random force procedure, and we shall discuss there how the rules referred to are established. In the meantime we concern ourselves in a very elementary way with the mathematical problems raised by this

procedure. It is necessary to discuss in what forms we can give the requisite statistical properties of the fluctuating forces, and how we can derive from them the desired statistical properties of the fluctuations of the macroscopic coordinates of the system.

The discussion which follows is similar to those given by Rice (1944) and by Lawson and Uhlenbeck (1950). For fuller accounts of the random variable concept, which is basic to the theory, reference may be made to Uspensky (1937) and Feller (1950) : the theory of random functions has been reviewed at an advanced level by Moyal (1949).

3.2. *Power Spectrum*

We have already seen that a randomly fluctuating quantity $y(t)$ can have associated with it various statistical properties such as a mean square value $\overline{y^2(t)}$, termed the power of the fluctuations, and a correlation function $\overline{y(t)y(t+\tau)}$. These averages must, in general, be supposed to be taken over the fluctuations of an ensemble of macroscopically identical systems. For the fluctuating deflection of a suspended system or for the fluctuating couple acting on the system the values of all such quantities are independent of t if the zero of time is chosen arbitrarily (i.e. independently of what the value of the fluctuation at the zero chosen happens to be). Random functions with this property and the processes generating them are said to be stationary. Clearly the power and the correlation function would depend on t if we were considering the fluctuations of an amplifier which was heating up or if, with a suspended system, we chose the zero of time in each experiment to be an instant at which the fluctuation in the deflection had some definite value: such non-stationary processes will not concern us here. For stationary processes we can consider, in addition to the ensemble averages already mentioned, time averages taken over the fluctuations of a single system. We assume that if the time average of a quantity is taken over a sufficiently long time it will agree with the ensemble average (the ergodic hypothesis of statistical mechanics).

We wish now to introduce a further very important statistical property characterizing a random function, viz. its power spectrum. In order to define the power spectrum let us suppose that we take a length t_0 of the fluctuations $y(t)$ of a particular instrument and, taking the arbitrary initial point of this length to be the zero of time, express $y(t)$ as a Fourier series:

$$y(t) = \sum_{r=1}^{\infty} (a_r \cos 2\pi rt/t_0 + b_r \sin 2\pi rt/t_0) \quad (0<t<t_0). \quad \ldots\ldots(40)$$

The a_r and b_r are constants which will have definite values for the length of record chosen. The absence of the constant term a_0 implies that the average of $y(t)$ over the interval is negligible as will be the case for the functions considered if t_0 is long. We now take a small frequency range f to $f+\delta f$ and consider the function $y_{\delta f}(t)$ which is obtained by suppressing all components in the expansion (40) of $y(t)$ except those for which the frequency r/t_0 lies in this range. It is assumed that in the limit in which t_0 goes to infinity the average of $y^2_{\delta f}(t)$ over t_0 will assume the same value for fluctuations of any of a set of macroscopically identical instruments and this limiting value is said to be the power of the random function in the frequency range considered. If the power in the small frequency range f to $f+\delta f$ is taken to be $G(f)\delta f$ then $G(f)$ is the power spectrum.

Two main results concerning $G(f)$ will be established.

(1) The power spectrum $G(f)$ is the Fourier cosine transform of the correlation function, i.e.

$$\overline{y(t)y(t+\tau)} = \int_0^\infty G(f)\cos 2\pi f\tau\, df \qquad \ldots\ldots(41\,a)$$

$$G(f) = 4\int_0^\infty \overline{y(t)\,y(t+\tau)}\cos 2\pi f\tau\, d\tau. \qquad \ldots\ldots(41\,b)$$

Thus knowing the power spectrum $G(f)$ is precisely equivalent to knowing the correlation function, and conversely. The relations (41 *a*), (41 *b*) are termed the Wiener–Khintchine relations and are of great importance in fluctuation theory.

(2) If a random force is acting on a linear system we can always formally regard it as acting at the input of an amplifier of which the output is the quantity of which the fluctuations are of interest. We can then associate with the system an impedance function $Z(f)$ such that an input force $\mathscr{R}A\exp(2\pi ift)$ produces an output $\mathscr{R}Z(f)A\exp(2\pi ift)$. Then if the power spectrum $G_I(f)$ of the input fluctuating force is known the power spectrum $G_O(f)$ of the output fluctuations is given by

$$G_O(f) = |Z(f)|^2 G_I(f). \qquad \ldots\ldots(42)$$

The importance of the power spectrum as an alternative to the correlation function comes largely from the simplicity of this relation.

To establish the above results we first note that it follows at once from (40) that for τ greater than zero

$$\frac{1}{t_0}\int_0^{t_0} y(t)y(t+\tau)\,dt = \tfrac{1}{2}\Sigma(a_r^2+b_r^2)\cos 2\pi r\tau/t_0, \qquad \ldots\ldots(43)$$

where, because of the artificial periodicity imposed by the Fourier series, $y(t+\tau)$ for $t_0<t+\tau\leqslant 2t_0$ is to be interpreted as $y(t+\tau-t_0)$, etc. Provided τ is negligible compared with t_0, so that this complication is of no consequence, the left-hand side is just an average of $y(t)y(t+\tau)$ over the time t_0, and since the random function is stationary the ensemble average of this time average will clearly be $\overline{y(t)y(t+\tau)}$. Consequently on taking the ensemble average of (43) we have

$$\overline{y(t)y(t+\tau)} = \tfrac{1}{2}\sum_{r=1}^{\infty}\overline{a_r^2+b_r^2}\cos 2\pi r\tau/t_0. \qquad \ldots\ldots(44)$$

If in place of $y(t)$ we considered $y_{\delta f}(t)$ the equation corresponding to (44) with τ equated to zero gives, for large t_0, an expression for the power in the frequency range f to $f+\delta f$, viz.

$$G(f)\delta f = \tfrac{1}{2}\sum_{f<r/t_0<f+\delta f}\overline{a_r^2+b_r^2}. \qquad \ldots\ldots(45)$$

Since t_0 is large the $\overline{a_r^2+b_r^2}$ may be assumed to vary slowly with r, and the terms on the right-hand side, $t_0\delta f$ in number since t_0 is large, may be taken to be equal. Consequently on writing f_r for r/t_0 we find that for large t_0

$$G(f_r) = \tfrac{1}{2}t_0(\overline{a_r^2+b_r^2}). \qquad \ldots\ldots(46)$$

Substituting for $\overline{a_r^2+b_r^2}$ in (44), writing $f_{r+1}-f_r$ for $1/t_0$, and letting t_0 go to infinity, we obtain

$$\overline{y(t)y(t+\tau)} = \lim_{f_{r+1}-f_r\to 0}\sum_{r=1}^{\infty} G(f_r)\cos 2\pi f_r\tau(f_{r+1}-f_r) = \int_0^\infty G(f)\cos 2\pi f\tau\, df.$$

The converse Wiener–Khintchine relation follows at once from the theory of the Fourier cosine transform.

The result (42) relating output to input power spectrum follows from (45) together with the elementary result $a_{0,r}^2+b_{0,r}^2=|Z(f)|^2(a_{I,r}^2+b_{I,r}^2)$ in which suffices 0 and I are again used to denote output and input quantities respectively.

3.3. *Experimental Determination of Power Spectrum of a Random Force*

It will be useful to collect together some results much used in subsequent calculations with power spectra and to indicate how the power spectrum of a random force can be obtained in practice.

Putting $\tau=0$ in (41 *a*) gives

$$\overline{y^2}=\int_0^\infty G(f)\,df. \qquad \ldots\ldots(47)$$

Combining this with (42) gives the much used result

$$\overline{y_0^2}=\int_0^\infty |Z(f)|^2 G_I(f)\,df. \qquad \ldots\ldots(48)$$

This enables us to obtain the mean square response of a linear instrument of known impedance function to a fluctuating force with a known power spectrum.

If $Z(f)$ is of appreciable magnitude only for frequencies near f_0 then (48) may be written

$$\overline{y_0^2}=G_I(f_0)\int_0^\infty |Z(f)|^2\,df. \qquad \ldots\ldots(49)$$

If a random force of unknown power spectrum is applied to the input of such a system, $G(f_0)$ can be determined by observing the mean square output, $\int_0^\infty |Z(f)|^2 df$ being known. By a series of such experiments covering a range of values of f_0, $G(f)$ can be determined experimentally. The time t_0 over which the output must be averaged must be long compared with the reciprocal of the bandwidth of the amplifier, since t_0 must be much larger than the time over which correlation in the output is appreciable and, by the well-known relation between the length of a wave packet and the range of frequencies composing it, this correlation time will be of the order of the reciprocal of the bandwidth. An alternative way of regarding this result is to note that according to (46), knowing $G(f_r)$ is equivalent to knowing the ensemble average of $a_r^2+b_r^2$: in the present method we are clearly estimating this ensemble average by averaging for a single record the $a_s^2+b_s^2$ for a range of s round r: the condition that t_0 should be much greater than the reciprocal of the bandwidth ensures that there will be a large number of components in the range of frequencies over which we are averaging, viz. the bandwidth.

If we require to estimate the power spectrum of a random function from a record of a particular realized function over a long time t_0, much greater than the time over which correlation is appreciable, we can determine the coefficients in a Fourier series analysis and use the fact that $G(f)\delta f$ is half the sum of the squares of the amplitudes of the components in the frequency range $\delta f(\geqslant 1/t_0)$. Alternatively we may estimate the correlation function as a time average and then use the Wiener–Khintchine relation (41 *b*). In this case we should estimate the correlation function from the record only for values of τ up to some value much less than t_0, and replace it by zero for larger values, for which the fractional

error in an estimate from the record would be large. It is interesting to verify that this method, like the first, is equivalent to finding an average of the squares of the components in a small frequency range: similar considerations are presented by Woodward (1948), although he bases his discussion on the more usual Fourier integral representation of the function in the range considered. The relative merits of estimating the power spectrum directly and deriving it from the correlation function are discussed by Davenport, Johnson and Middleton (1952).

3.4. *Probability Distributions*

The statistical properties which we have so far used and which will be sufficient for most of the subsequent discussion are of the nature of averages and do not by any means characterize completely the statistical properties of a random function (see Woodward 1948). For some purposes it is necessary to consider various probability distributions, e.g. the probability that the values of the function at a prescribed set of instants will lie in specified ranges. It is frequently convenient to consider the probability distributions of the a_r and b_r. For most of the random forces with which we are concerned here these random variables are all independent and normally distributed about zero. Random functions of this nature are said to be gaussian random, and it can be shown that if the input of a linear system is gaussian so also will be the output. These topics are much more fully discussed by Rice (1944, 1945), Lawson and Uhlenbeck (1950), Wang and Uhlenbeck (1945).

3.5. *Fluctuations in Circuit with Inductance and Resistance*

3.5.1. *Power spectrum method.*

We shall now apply the above theory to discuss the fluctuations of current in a circuit consisting of an inductance L and a resistance R. In this particular case the methods which we used in §2 are available and would actually be simpler to apply, but it is interesting to compare the two approaches. We again suppose that the fluctuations are caused by a rapidly fluctuating 'force' (in this case a voltage). The fact that the fluctuations are rapid can be embodied in the statement that the correlation function for the random voltage $\overline{E(t)E(t+\tau)}$ drops practically to zero in a time τ' (say) much shorter than the time constant L/R : $\tau' \sim 10^{-13}$ sec (the collision time of the electrons). It follows that for any frequency low enough to produce an appreciable response in the circuit, and for values of τ for which the correlation function for the voltage is appreciable, $f\tau$ will be negligible. Consequently in using (41 *b*) to obtain the power spectrum from the correlation function, we can, so long as we confine ourselves to these frequencies, replace $\cos 2\pi f\tau$ by unity. This means that for these frequencies, i.e. for frequencies not enormously high, the power spectrum $G_{\mathrm{E}}(f)$ will be constant. In fact

$$G_{\mathrm{E}}(f) = 4\int_0^\infty \overline{E(t)E(t+\tau)}\,dt = 2\overline{EE} \quad (f \ll 1/\tau') \qquad \ldots\ldots(50)$$

where

$$\overline{EE} = \int_{-\infty}^{+\infty} \overline{E(t)E(t+\tau)}\,d\tau. \qquad \ldots\ldots(51)$$

Now a voltage $\exp(2\pi ift)$ in the circuit will produce a current $\exp(2\pi ift)[R+2\pi ifL]^{-1}$ so that $[R+2\pi ifL]^{-1}$ is the impedance function $Z(f)$.

It follows from (42) and (50) that the power spectrum of the current $G_i(f)$ (the output power spectrum in the present case) is given by

$$G_i(f) = 2\overline{EE}/[R^2 + 4\pi^2 f^2 L^2]. \quad \ldots\ldots(52)$$

The correlation function for the current is obtained by using (41 *a*) which gives

$$\overline{i(t)i(t+\tau)} = \int_0^\infty \frac{2\overline{EE}\cos 2\pi f\tau}{R^2 + 4\pi^2 L^2 f^2}\,df = \frac{\overline{EE}}{2RL}\exp(-R|\tau|/L). \quad \ldots\ldots(53)$$

This would give the required information about the fluctuation if we knew $\overline{EE}$. In this case, however, we can argue in reverse and deduce the value of $\overline{EE}$ from the equipartition result $\frac{1}{2}L\overline{i^2} = \frac{1}{2}kT$. This means that the fluctuating force must be such that

$$\overline{EE} = 2Rk\mathbf{T}. \quad \ldots\ldots(54)$$

The power spectrum of the random voltage is therefore, according to (50), given by

$$G_E(f)\delta f = 4Rk\mathbf{T}\delta f. \quad \ldots\ldots(55)$$

This is in fact Nyquist's famous formula for the power spectrum of the random voltage which has to be associated with a resistance. It will be dealt with in the following section (§4.3).

3.5.2. *Correlation function method.*

In the foregoing example we derived the correlation function of the output from the correlation function of the input by using the corresponding power spectra as intermediaries in the discussion. It is possible to proceed more directly. In any particular experiment the fluctuating voltage $E(t)$ and the fluctuating current $i(t)$ are related by the equation

$$L\frac{di}{dt} + Ri = E(t). \quad \ldots\ldots(56)$$

If the value of the current at the arbitrarily chosen zero of time is i_0, then we can give this relation in the form

$$i = i_0 e^{-Rt/L} + \frac{1}{L}e^{-Rt/L}\int_0^t e^{Rx/L}E(x)\,dx. \quad \ldots\ldots(57)$$

We can neglect the first term if we confine our interest to the solution for large t. It then follows that in this particular experiment, for large t and any τ,

$$\begin{aligned}\overline{i(t)i(t+|\tau|)} &= \frac{1}{L^2}e^{-R(2t+|\tau|)/L}\int_0^t\int_0^{t+|\tau|} e^{R(x+y)/L}\overline{E(x)E(y)}\,dy\,dx \\ &= \frac{1}{L^2}e^{-R(2t+|\tau|)/L}\int_0^t\int_{-x}^{t+|\tau|-x} e^{R(2x+\delta)/L}\overline{E(x)E(x+\delta)}\,d\delta\,dx,\end{aligned}$$

where $y = x+\delta$. We can now average over an ensemble of experiments and get

$$\overline{i(t)i(t+|\tau|)} = \frac{1}{L^2}e^{-R(2t+|\tau|)/L}\int_0^t e^{2Rx/L}\left\{\int_{-x}^{t+|\tau|-x} e^{R\delta/L}\,\overline{E(x)E(x+\delta)}\,d\delta\right\}dx \;(t \gg L/R). \quad \ldots\ldots(58)$$

This relates the correlation functions of input and output. With the form already assumed for the correlation function of the voltage this is easily evaluated. Since $e^{R\delta/L}$ will be practically unity for all values of δ for which $\overline{E(x)E(x+\delta)}$ is appreciable it can be replaced by unity in the inner integral. Moreover, the limits of this integrand, which always lie on either side of zero, can be extended to infinity without appreciable error; the inner integral thus becomes just $\overline{EE}$.

The whole right-hand side can then be evaluated for $t \to \infty$, and we find

$$\overline{i(t)i(t+\tau)} = \overline{i(t)i(t+|\tau|)} = \frac{\overline{EE}}{2RL} e^{-R|\tau|/L}. \qquad \ldots\ldots(59)$$

This reproduces the previous result. The procedure just given is much used in the literature in cases in which the correlation function of the random force concerned has the specially simple form assumed here.

3.6. *Time Integrals of Current and Voltage*

Let us consider the mean square value of the net charge which current fluctuations will cause to pass in a time S round a circuit with inductance and resistance. The charge Q is just the integral of the current over the interval S: consequently

$$\overline{\Delta Q^2} = \overline{\left\{\int_0^S i\,dt\right\}^2} = \int_0^S \int_0^S \overline{i(t_1)i(t_2)}\,dt_1\,dt_2. \qquad \ldots\ldots(60)$$

If we replace the variables t_1 and t_2 by t_1 and $t_1+\tau$ and use the expression (53) for the correlation function of the current we get

$$\overline{\Delta Q^2} = \frac{2kT}{R}\left\{S - \frac{L}{R}(1 - e^{-RS/L})\right\}. \qquad \ldots\ldots(61)$$

When S is much greater than the time constant of the circuit L/R this gives Einstein's result $\overline{\Delta Q^2} = 2kTS/R$. It should be noted, however, that if S is less than L/R the value of $\overline{\Delta Q^2}$ may be much less than that given by Einstein's result.

On the other hand the mean square value of the integral of the fluctuating voltage over a time S is given by a result analogous to Einstein's which, because of the negligible correlation of the voltage fluctuations, holds for all macroscopic S. This is why we chose in §2.1.2 to establish in the first place a limit for the accuracy in measuring a voltage rather than to consider directly the accuracy in measuring a current. The result, already used in §2.1.2 is easily established. We get

$$\overline{\left\{\frac{1}{S}\int_0^S E(t)\,dt\right\}^2} = \frac{1}{S^2}\int_0^S \int_0^S \overline{E(t_1)E(t_2)}\,dt_1\,dt_2$$

$$= \frac{1}{S^2}\int_0^S \int_{-t_1}^{S-t_1} \overline{E(t_1)E(t_1+\tau)}\,d\tau\,dt_1. \qquad \ldots\ldots(62)$$

Since the time over which correlation is appreciable is negligible $\overline{E(t_1)E(t_1+\tau)}$ will be zero for values of τ appreciably different from zero and the limits of the inner integral can be extended to infinity; this integral then becomes just $\overline{EE}$ and we get the important result (applied later to a variety of fluctuations)

$$\overline{\left\{\frac{1}{S}\int_0^S E(t)\,dt\right\}^2} = \overline{EE}/S = G_E(0)/2S. \qquad \ldots\ldots(63)$$

The alternative expression in terms of the power spectrum of the voltage follows from the Wiener–Khintchine relation (41·*b*).

The fluctuating voltage is that associated with a resistance R; consequently $\overline{EE}$ is given by (54) and the right-hand side of (63) is just $2RkT/S$. This is the minimum uncertainty in measuring a voltage in time S as explained in §2.1.2 and the corresponding result for the minimum uncertainty in measuring a current follows, viz.

$$\overline{\Delta i^2} = 2kT/RS. \qquad \ldots\ldots(64)$$

3.7. *Measurement of a Current with an Ideal Amplifier*

It is of interest to consider the measurement of a current in a circuit of resistance R by making a single reading of the output of an amplifier which is ideal in the sense that the only fluctuating force which has to be considered is the fluctuating voltage in the circuit. The amplifier will be described sufficiently by giving the impedance function relating the output θ to the input voltage. Since the random voltage has a practically constant power spectrum we get from (48) and (55)

$$\overline{\Delta\theta^2} = 4Rk\mathfrak{T}\int_0^\infty |Z(f)|^2\,df. \qquad \ldots\ldots(65)$$

The estimate of the current will be obtained from the relation $\theta = Z(0)Ri$ and the mean square fluctuation in the estimate of the current will be given by

$$\overline{\Delta i^2} = \frac{1}{|Z(0)|^2R^2}\overline{\Delta\theta^2} = \frac{4k\mathfrak{T}}{R}\frac{1}{|Z(0)|^2}\int_0^\infty |Z(f)|^2\,df. \qquad \ldots\ldots(66)$$

If an amplifier of impedance function $Z(f)$ is being used to measure a sinusoidally varying quantity with frequency f_0 the equivalent noise bandwidth of the amplifier is defined as

$$\left\{\int_0^\infty |Z(f)|^2\,df\right\}\Big/|Z(f_0)|^2$$

and is written B. The significance of the term 'bandwidth' is that the mean square response of the amplifier to an input noise with a power spectrum which is effectively constant in frequency will be the same as that of a sharp cut-off amplifier with bandwidth B and an impedance function which has the constant value $Z(f_0)$ throughout its passing band. In the present instance f_0 is zero and we can rewrite (66)

$$\overline{\Delta i^2} = 4k\mathfrak{T}B/R. \qquad \ldots\ldots(67)$$

Comparing this with (64) we see that taking the average over a time S is a process with a bandwidth $1/2S$, a result sometimes quoted in the literature. This can also be established directly from the definition of B.

It is easy to see that if a linear instrument is being used to measure a sinusoidal current of frequency f_0, say, so that B is taken about f_0, the right-hand side of (67) is still significant and indicates the mean square value of the current of frequency f_0 which will give an output with the same mean square value as the noise output. It is therefore conventional to take it as giving the mean square value of the minimum detectable signal.

It is interesting to note that a galvanometer becomes an ideal amplifier if is made sufficiently large: this follows from the fact that when we combine it with an averaging process we attain the minimum possible uncertainty, which would not be the case if, for example, the fluctuating couple were important. Increasing G has the effect of increasing both the response to the current and the response to the fluctuating voltage, relative to the response to the fluctuating couple. If G is not large enough to make the fluctuating couple negligible, we may say that the coupling of the amplifier to the quantity to be measured and to the fluctuations inevitably associated with it is not large enough.

3.8. *Fluctuations of a Torsionless Suspension*

In the first section we treated the torsionless suspension as a limiting case. We can, however, start from the equation

$$I\,d\omega/dt+\kappa\omega=P(t)$$

where $P(t)$ is a fluctuating couple. This is analogous to the equation $L\,di/dt+Ri=E(t)$ and the statistical mechanical result $\frac{1}{2}I\overline{\omega^2}=\frac{1}{2}kT$ is analogous to $\frac{1}{2}L\overline{i^2}=\frac{1}{2}kT$. If we assume that $P(t)$, like $E(t)$, will have a correlation function which drops to zero in a negligibly short time, we get by analogy with eqns. (54), (53) and (61)

$$\overline{PP}=2\kappa kT, \qquad \ldots\ldots(68)$$

$$\overline{\Delta\omega(t)\Delta\omega(t+\tau)}=\frac{kT}{I}\exp\left(-\frac{\kappa|\tau|}{I}\right), \qquad \ldots\ldots(69)$$

$$\overline{\Delta\theta^2}=\overline{\left\{\int_0^S\Delta\omega(t)\,dt\right\}^2}=\frac{2kT}{\kappa}\left[S-\frac{I}{\kappa}\left\{1-\exp\left(-\frac{\kappa}{I}S\right)\right\}\right]. \qquad \ldots\ldots(70)$$

The last result was used in § 2 (eqn. (24)).

By (68) and the analogue of (63) the mean square error of the best estimate which can be made in a time S of a couple acting on a suspended system with damping constant κ is $2\kappa kT/S$. The proof corresponds exactly to that used in § 3.6 to establish the minimum uncertainty in the estimate of a voltage in a circuit of resistance R.

3.9. *Further Mathematical Considerations*

The discussion which we have just given of the use of correlation functions and power spectra in treating the fluctuation of linear systems provides all the mathematical equipment necessary for the understanding of the following sections. It seems desirable, however, that we should indicate the sort of problem which calls for quite different mathematical methods, outline some of the more important of these methods, and give references to more comprehensive treatments.

The whole subject becomes much more complicated and difficult when non-linear systems (e.g. systems including a rectifying element) are considered. A very lucid introduction to the mathematical methods available under these circumstances is given by Lawson and Uhlenbeck in the chapter already mentioned. We shall confine ourselves to two very brief remarks indicating typical complications which arise. In the first place, calculating even the mean square magnitude of the output depends on knowing appropriate distribution functions for the input. Secondly, there is of course no question of a linear superposition of the response to the signal and to the random force separately: what matters therefore is the distribution function of the signal plus the random force. Rice discusses a number of the resulting problems. Further references will be found in the paper by Middleton (1951).

Other mathematical considerations present themselves when we turn our attention to the mechanisms giving rise to fluctuating forces. The discussion here could extend indefinitely into mathematical kinetic theory, but we shall confine ourselves to indicating how we can proceed when analysis of the physical model for the fluctuating force leads to a description of its statistical properties in a form different from those already mentioned. Thus, for example, a fluctuating voltage may be regarded as arising from a random series of pulses

of voltage. This description appears very naturally in analysing shot noise. The simplest instance of this is the fluctuation in the voltage across a load in series with a diode operated under saturation conditions: this is additional to the fluctuating force associated with a resistance, and has its origin in the fact that the current is made up of a series of independent current pulses, each one corresponding to the emission of an electron from the cathode. If the saturation condition is not imposed the situation is more complicated, because the current pulses are no longer independent: the space charge gives rise to a barrier which the electrons must surmount and the height of this barrier depends to some extent on the number of electrons which have just passed over it. This results in a net reduction of the mean square fluctuation in the voltage.

With all the alternative descriptions of random forces either we may aim at deriving directly a corresponding description of the fluctuations in the output, or we may try to turn the description at once into a description in terms of the correlation function or power spectrum and then proceed in the usual way. When the random force is described as a random series of pulses either method is practicable. Thus, to use the first, the direct method, we simply calculate the response of the linear system to one of the pulses which make up the input; the output will then be just a random series of such responses. The conversion of the pulse description into terms of a correlation function description is effected by means of a theorem due to Campbell and some extensions of it, which will now be stated briefly.

We suppose that the shape of the pulse is given by $f(t)$ (the zero of time being chosen arbitrarily relative to the pulse) and that there are on the average λ such pulses in unit time. Let the resulting random function be denoted by y and the random function obtained from it by subtracting the average value by y_m. The correlation function for y_m is denoted by $\psi_{y_m}(\tau)$. Then

$$\bar{y} = \lambda \int_{-\infty}^{+\infty} f(t)\, dt, \qquad \ldots\ldots(71)$$

$$\overline{y_m^2} = \lambda \int_{-\infty}^{+\infty} f^2(t)\, dt, \qquad \ldots\ldots(72)$$

$$\psi_{y_m}(\tau) = \lambda \int_{-\infty}^{+\infty} f(t) f(t+\tau) dt. \qquad \ldots\ldots(73)$$

By combining (73) with (41 *b*) we find that if

$$f(t) = \int_{-\infty}^{+\infty} S(f)\, e^{2\pi i f t} df$$

then $G_{y_m}(f)$, the power spectrum of y_m, is given by

$$G_{y_m}(f) = 2\lambda |S(f)|^2. \qquad \ldots\ldots(74)$$

These results can be applied to either the input or the output functions. A general discussion of the results related to determining the mean values of higher power is given by Rice (1944, 1945). Rice is concerned with the distribution functions for noise described in terms of pulses and shows that the random functions tend to become gaussian as the frequency of the pulses grows. An extensive discussion of problems of this nature is given in the paper by Middleton (1951) already mentioned.

Another device which is much used in the literature of the subject is to assume that there exists a time interval Δt which is short enough to ensure that

the change during that interval in the coordinates of the system on which the random force acts will be small, but long enough to make the value of the random force at the end of the interval independent of the value at the beginning. In effect, the random force is replaced by a sequence of equally and closely spaced impulses of varying magnitude. The mean square value of the impulse is closely related to the integral of the correlation function of the random force, being just $\overline{FF}\Delta t$, and many of the arguments given in the early literature which involve this 'physically infinitely small' time interval could easily be rephrased in terms of a correlation function: this applies to arguments in which the requisite properties of the pulse in the interval are determined by the requirement that the fluctuations of the system on which it acts should conform to statistical mechanical results. On the other hand the device is useful when determining the properties of the force from a physical model: an example is provided by the analysis of the fluctuation of a suspended system given by Uhlenbeck and Goudsmit (1929). Here the mean square impulsive moment imparted by the gas molecules in a time Δt is calculated from the kinetic theory of gases. Since the mean square impulsive moment is $\overline{PP}\Delta t$ we can determine $\overline{PP}$ from it and so the value of the constant power spectrum of the couple.

The physically infinitely small time interval also makes its appearance in the diffusion theory of Brownian motion. We shall merely sketch this approach. The method is used in a particularly cogent way by Kramers (1940) and Chandrasekhar (1943) and is discussed in an illuminating manner by Wang and Uhlenbeck (1945). In this method we first calculate the probability distribution of the change in momenta in a time Δt of a system with given initial coordinates and momenta. It is then possible to set up with the aid of this information an equation giving the way in which the distribution of an assembly of such systems changes with time. The time-independent solution of this equation gives the equilibrium probability distribution. This method is perhaps rather remote from the sort of fluctuation considerations which are our main concern. However it has the interest that it is not confined in application to linear systems. The diffusion of a large number of non-interacting particles suspended in a liquid can be treated in this way, and indeed Einstein gave a discussion of Brownian motion of a suspended particle which was closely related to the ordinary diffusion equation. As a matter of fact, however, this ordinary diffusion equation is not quite the same as the equation giving the change of the probability distribution with time mentioned above, but is an approximation to it valid when we are not interested in what happens in times over which there is appreciable correlation in velocity. The equation referred to above applies to a distribution in position and momentum space, the diffusion equation applies to a distribution in ordinary space.

§4. PHYSICAL BASIS

Our programme now calls for some remarks on the physical basis of the theory. To a considerable extent the subject has advanced by virtue of a series of plausible assumptions which have been justified very convincingly by the success of predictions based on them. It certainly cannot be claimed that these assumptions have been derived rigorously, in all their generality, from the general body of statistical mechanics and kinetic theory. In particular instances detailed derivations have been given and these will be mentioned in due course. We shall however be content for the most part to state more or less dogmatically

the principles on which the usual applications of the theory rest, pointing out the physical nature of the assumptions involved, and indicating where possible the sort of arguments which serve to render these assumptions plausible.

4.1. *Equilibrium Systems*: *Equipartition*

The discussion of the first section was based on an application of the equipartition principle. This asserts that if a coordinate q or a momentum p of an equilibrium system appears in the energy (or more strictly the Hamiltonian) only in a square term (i.e. as aq^2 or bp^2 with a and b positive constants), then the average of that term taken over an ensemble of systems is $\frac{1}{2}kT$. A discussion generalization of this result is given by Tolman (1938).

In applying the equipartition theorem to a suspended system we effectively assumed the Hamiltonian of this system to be

$$H=\tfrac{1}{2}c\theta^2+\frac{1}{2I}p_\theta^2 \qquad \text{.....(75}$$

where $p_\theta=Id\theta/dt$ the momentum conjugate to θ. This, however, neglects interaction with the molecules of the medium surrounding the suspended system (the main mechanism of damping) and with the molecules of the suspended system itself (internal friction). Now it is always assumed in statistical mechanics that there are interactions which are not included in the Hamiltonians with which we calculate. These absent terms are regarded as allowing the energy exchange necessary for setting up equilibrium, but as being, for all likely configurations, negligibly small compared with the terms included in the Hamiltonian. We have assumed that in the case of the suspended system the interaction which gives rise to macroscopic damping is of this type.

That the interaction energy is small and that consequently the equipartition results hold is plausible, except possibly in the case of large damping. An argument due to Einstein (1906), (see also Fürth 1948), shows very convincingly that the result $\frac{1}{2}c\overline{\theta^2}=\frac{1}{2}kT$ holds independently of what the damping may be. If all the microscopic coordinates are measured relative to a mean plane in the suspended system of which the position is given by θ, then by symmetry the Hamiltonian will, in the absence of torsion, be independent of θ. Introducing torsion simply adds a term $\frac{1}{2}c\theta^2$ to the Hamiltonian, it being assumed that interaction with the microscopic coordinates of the suspension fibre itself may be neglected. It follows at once from our statement of the equipartition principle that $\frac{1}{2}c\theta^2$ is just $\frac{1}{2}kT$.

This argument does not justify the result $\frac{1}{2}\overline{p_\theta^2}/I=\frac{1}{2}kT$ since the interaction term will certainly depend to some extent on the momentum p_θ. This dependence may give rise to macroscopic effects other than damping, as when, for example, the suspended system drags an appreciable amount of the surrounding medium with it, in which case the interaction term would not be negligible. It is usual to assume, however, that in the absence of macroscopic effects other than damping, the interaction term will be negligibly small and that consequently the simple equipartition result for the momentum will hold.

The experimental verification of the mean potential energy result has, as already mentioned, been carried out by Kappler (1931). The experimental verification of the equipartition result for the kinetic energy is not quite so direct, since what is measured is not in the first instance an instantaneous angular velocity

but an average of the angular velocity over an interval of time which may or may not be small enough for this average to be effectively the instantaneous velocity. For some remarks on this see MacDonald (1950) and Fürth (1950). If for simplicity we assume the suspension torsionless we can use (24) to get an expression for the mean square value of the average of the angular velocity over a time τ, viz.

$$\frac{\overline{\Delta\theta^2}}{\tau^2} = \frac{2\boldsymbol{k}T}{\kappa\tau^2}\left[\tau - \frac{I}{\kappa}\left\{1 - \exp\left(-\frac{\kappa}{I}\tau\right)\right\}\right]. \qquad \ldots\ldots(76)$$

It will be seen that if τ is much less than I/κ the right-hand side will have practically its limiting value $\boldsymbol{k}T/I$. Kappler (1938) achieved conditions under which an observation time less than I/κ was feasible by working in a rarified gas (so as to make κ small). He was able to verify the equipartition result for the kinetic energy very satisfactorily under these circumstances. Such a direct verification seems hardly possible with large damping. It would then be necessary to work with the general form of (76) applicable for any τ. The limiting form for very large τ has been checked by Perrin (1909) for the case of particles suspended in a liquid.

Similar considerations apply to a circuit with inductance, capacity and resistance. For this, as for the suspended system, the potential energy result $\frac{1}{2}\overline{Q^2}/C = \frac{1}{2}\boldsymbol{k}T$ (Q is the charge on the condenser of capacity C) is justified by Einstein's argument. The kinetic energy result $\frac{1}{2}L\overline{i^2} = \frac{1}{2}\boldsymbol{k}T$ requires the assumption that the energy of interaction between the current and the microscopic coordinates of the conductor is small. Numerous verifications of these results are available (Johnson 1928, Moullin 1938). Only in recent Russian work have departures been found which are regarded as significant. These anomalies arise with resistances made of alloys (Pumper 1949). The results are attributed to non-equilibrium processes in the material of the resistance. Possible examples of such non-equilibrium processes would be recrystallization of the material, or the release of strains. Another theoretical possibility would be radioactive disintegration of some of the nuclei present.

The fluctuations of a galvanometer are of special interest, as they are to be regarded as arising from both molecular bombardment of the suspension and electrical fluctuations in the circuit. Experimental verifications for this case have been mentioned in §4.

Leaving aside now the question of detailed justification, we may state baldly the procedures used in applying what may be called the statistical mechanics of macroscopic variables to macroscopic systems. We ignore all seats of energy dissipation (electrical resistance, damping, etc.), write down the Hamiltonian of the resulting system in so far as it is described by macroscopic variables and then apply the statistical mechanical results applicable to a system in small interaction with its surroundings. This, as already indicated, is the procedure implicit in applications of the equipartition principle to simple macroscopic systems. It has been applied to more complicated systems. Barnes and Silverman (1934) discuss a number of cases and the method has been applied to a galvanometer with appreciable inductance in its circuit by Jones and McCombie (1952) and by Smith (1951).

There is always, of course, a certain arbitrariness in setting up the Hamiltonians as to what one takes to be a fluctuating coordinate and what one takes to be a fixed parameter. Thus in the theory of the galvanometer it is usual to regard

the strength of the magnetic field as a fixed parameter. However there will doubtless be fluctuations in the position of the domain boundaries in the magnet which will give rise to fluctuations in the strength of the magnetic field. It is not difficult to show that working with the wider system in which these fluctuations are allowed does not lead to a different estimate of $\frac{1}{2}c\overline{\theta^2}$, $\frac{1}{2}I\overline{(d\theta/dt)^2}$ and $\frac{1}{2}L\overline{i^2}$.

4.2. *Equilibrium Systems: Random Forces*

To be definite, let us consider once again the fluctuations of a suspended system. As we have just seen, some features of these fluctuations follow from statistical mechanical considerations. An alternative approach is, however, available, for we can regard the fluctuations as arising from bombardment of the suspended system by molecules of the surrounding medium, and it is clearly possible, at least in principle, to obtain any required information about the fluctuations from a kinetic theory analysis of this process. In the particular case in which the surrounding medium is a rarified gas, such an analysis has, as already mentioned, been carried through by Uhlenbeck and Goudsmit (1929).

Comparing the two approaches we may say that the application of kinetic theory consists in discussing in detail the small interaction with the surroundings, the mere existence of which is all that is required for the statistical mechanical arguments. The kinetic theory approach in general, and the analysis of Uhlenbeck and Goudsmit in particular, not only reproduce results which can be obtained by the statistical mechanical approach, but also gives other information not so obtainable, e.g. the power spectrum of the fluctuations. That the power spectrum cannot be derived from the statistical mechanical procedure just outlined is, of course, immediately evident from the fact that constants relating to the dissipative forces usually appear in the expression for it. We may, however, remark in passing that attempts are being made to apply statistical mechanical ideas to irreversible processes and one of the first steps is to endeavour to discuss the power spectrum of the fluctuations of equilibrium systems from a statistical mechanical point of view (cf. Callen and Welton 1951, Greene and Callen 1951 and Callen and Greene 1952).

Now we saw in § 2 that a few very simple and natural assumptions enabled us to supplement the statistical mechanical arguments sufficiently to obtain the correlation function of the fluctuations; this, by the Wiener–Khintchine relations, is tantamount to obtaining the power spectrum. It was sufficient to assume that interaction with the surroundings could be regarded as taking the form of two couples acting on the suspended system, the macroscopic damping couple proportional to the angular velocity but opposite in sense, and a rapidly fluctuating random couple.

In specific instances this assumption can be justified by kinetic theory arguments. In the Uhlenbeck and Goudsmit analysis the damping couple appears as the average, taken over an ensemble of systems with a given angular velocity, of the impulsive moment arising from the molecular bombardment in a physically infinitely small interval Δt; the random part is represented by the fluctuations in this impulsive moment. A detailed kinetic theory analysis of the forces exerted by their surroundings on various other systems has proved possible. Kirkwood (1946) has shown that, on certain plausible assumptions as to the behaviour of some integrals which emerge in his calculations, the forces acting on a molecule in a liquid can be split into a dissipative and a random part;

Chandrasekhar (1949) has reached similar conclusions in discussing a problem in stellar dynamics, and Greene (1951) has shown how Kirkwood's very general considerations apply in a specific case (a sphere acted on by a gas of non-interacting molecules).

The random force theory of fluctuations depends on assuming that we can always obtain an adequate model of the interaction with the surroundings by associating a random force having suitable statistical properties with each dissipative force acting on the system. It turns out that we can, at least for a wide range of linear systems, ascribe the statistical properties in such a way as to lead to fluctuations satisfying the requirements imposed by statistical mechanical considerations. Moreover these statistical properties depend only on the constant describing the dissipative force, Boltzmann's constant and the absolute temperature. It is not surprising from our point of view that the properties of the fluctuating force should depend on the magnitude of the dissipative force, since the fluctuating force is to be regarded as counteracting the tendency of the dissipative force to remove the fluctuation energy from the system.

It is usual to assume, as we do here, that the fluctuating force is quite uncorrelated with the values of the coordinates and momenta of the system on which it is acting. Provided the damping force is linear, i.e. proportional to the 'velocity' which calls it into play, it is possible even with this restriction to choose the properties of the random force in such a way as to reproduce the statistical mechanical results. It appears, however, from the paper by Kramers (1940) already mentioned that if the damping force is non-linear it may be necessary to assume a dependence of the properties of the fluctuating force on the corresponding velocity in order that the statistical mechanical results may be reproduced.

The rules for associating fluctuating forces with the various dissipative forces are established by considering particular systems and finding fluctuating forces which will produce fluctuations with the properties required by statistical mechanics. Thus we have already seen (§ 3.5) that the fluctuations in a circuit consisting of an inductance L and resistance R in series would conform to the statistical mechanical requirements if we associated with the resistance a random voltage appearing in series with it with a constant power spectrum $G_{\mathrm{E}}(f)$ given by

$$G_{\mathrm{E}}(f) = 4Rk\mathrm{T}. \qquad \ldots\ldots(77)$$

On considering a system suspended by a torsionless fibre, it followed in an exactly analogous way that fluctuations with the required properties would be obtained if we associated with the damping a random couple P with constant power spectrum $G_{\mathrm{P}}(f)$ given by

$$G_{\mathrm{P}}(f) = 4\kappa k\mathrm{T}, \qquad \ldots\ldots(78)$$

where κ is the damping constant.

It is of course desirable to check in a wide variety of situations that this association of random forces with resistances and seats of damping produces results conforming to statistical mechanical requirements. It can be shown (Lawson and Uhlenbeck 1950) that the use of (77) gives results correct in this sense when applied to an arbitrary network of inductances, capacities and resistances. Similarly the random force (78) is adequate when applied to a system for which the torsion constant is not zero. We may emphasize that setting up such random force models merely to reproduce statistical mechanical

results is, strictly speaking, tautological and unnecessary, although it may in particular instances allow us to arrive at the required result more quickly than would a direct application of statistical mechanics. The model provides us with new information only when we *assume* that it is adequate to give results which are not given directly by statistical mechanical arguments (e.g. the power spectrum of the fluctuations).

4.3. *Nyquist's Result*

In arriving at the power spectra of voltage and couple fluctuations given by (77) and (78), we assumed from the start, on physical grounds, that the correlation time of the fluctuating forces would be negligibly small and that consequently the power spectrum would be practically constant. It is of interest to enquire whether the shape of the power spectrum of the random forces can be determined purely by statistical mechanical considerations. This question has been discussed very extensively in the particular case of the random voltage associated with a resistance. In fact the first formulation of the result (77) was given by Nyquist (1928) on the basis of statistical mechanical arguments: the equivalent correlation function result (54) had been used earlier, but the precise relation between the two results was not then known, as the Wiener–Khintchine relations did not emerge till later (Wiener 1930, Khintchine 1934).

Several objections have been advanced against Nyquist's proof (Moullin 1938, Fürth 1948). Here we shall reframe it in such a way as to make explicit use of the procedure for applying statistical mechanics outlined at the end of §4.1 and shall employ a definition of the power spectrum which is essentially that of §3.2. Instead of using (40) to analyse the random voltage appearing in series with a resistance R we shall use the half range sine series

$$E(t)=\Sigma A_r \sin \pi r t/t_0 \qquad 0<t<t_0. \qquad \ldots\ldots(79)$$

By a discussion similar to that of §3.2 it follows that in the limit of large t_0

$$G(f_r')=t_0\overline{A_r^2} \qquad (f_r'=r/2t_0). \qquad \ldots\ldots(80)$$

We now suppose two equal resistances R introduced at the ends of a lossless transmission line of length d to which they are matched (i.e. $R=(L/C)^{1/2}$, where L and C are the inductance and capacity per unit length). Half the fluctuating voltage appearing in series with either resistance will be transmitted along the line as a wave train (of velocity $v=(LC)^{-1/2}$) which will be completely absorbed at the other resistance. The energy in what would be the rth normal mode if resistance were absent (i.e. if the ends were shorted) is that in the rth component of a Fourier sine series analysis of the voltage in the line. Now each of the two independent wave trains will contribute $\frac{1}{4}A_r^2$ to the mean square amplitude of this component A_r being calculated for $t_0=d/v$. The potential energy in the mode will therefore be $\frac{1}{4}Cd\frac{1}{2}\overline{A_r^2}$ and this must equal $\frac{1}{2}kT$. It follows that $t_0\overline{A_r^2}$ is equal to $4RkT$ and since this holds for arbitrarily large t_0 we get, on substituting in (80),

$$G(f_r')=4RkT. \qquad \ldots\ldots(81)$$

Since for any f we can find a t_0, greater than any given value, for which there will be an f_r' equal to f, we may replace $G(f_r')$ by $G(f)$ and so arrive at Nyquist's result.

An alternative statistical mechanical approach suggests itself. Instead of trying to determine the whole power spectrum of the voltage fluctuations

associated with a resistance by discussing the fluctuations of a system in which it is included, we may include it in a system of which the fluctuations depend only on the value of the power spectrum over a small range of frequencies, for instance, we may imagine the resistance connected in series with an inductance L and capacity C. If, keeping the resonant frequency $1/2\pi(LC)^{1/2}$ fixed at f_0, we let L become very large, we can make the circuit very sharply resonant, so that only the components of the fluctuating voltage in a small region round f_0 will produce an appreciable response. The value of $G(f_0)$ must then be such as to make $\frac{1}{2}L\overline{i^2}=\frac{1}{2}kT$.

The transfer impedance from voltage to current (i.e. the admittance) is given by

$$Z(f)=\{R+2\pi i(Lf-1/4\pi^2Cf)\}^{-1}. \qquad \ldots\ldots(82)$$

If f_0^2 is written for $1/4\pi^2LC$, we obtain

$$|Z(f)|^2=\{R^2+4\pi^2L^2(f-f_0^2/f)^2\}^{-1}. \qquad \ldots\ldots(83)$$

According to (48) $\overline{i^2}$ is given by

$$\int_0^\infty G_E(f)\,|Z(f)|^2\,df.$$

In the limit only the value of $G_E(f)$ at $f=f_0$ is important and we may write

$$\overline{i^2}=G_E(f_0)\int_0^\infty \frac{df}{R^2+4\pi^2L^2(f-f_0^2/f)^2}. \qquad \ldots\ldots(84)$$

The value of the integral is independent of f_0^2 and is in fact always equal to $1/4LR$. Consequently $L\overline{i^2}$ is just $G_E(f_0)/4R$ and since this must be equal to kT we get

$$G_E(f_0)=4RkT. \qquad \ldots\ldots(85)$$

This proof is very simple and useful for didactic purposes; the hypothetical procedure involved conforms closely to the way in which one would measure the power spectrum experimentally. Van der Ziel (1952) argues similarly.

4.4. *Extension of Nyquist's Result to an Impedance*

If we form an impedance by connecting a network of inductances, capacities and resistances across two terminals each resistance will have a random voltage in series with it and, according to Thévenin's theorem, these will together be equivalent to a single fluctuating voltage in series with the impedance, with statistical properties independent of any load placed across the impedance. The power spectrum of this random voltage can at once be expressed in terms of the impedance, $\rho(f)=R(f)+iX(f)$, by an argument due to Nyquist. We imagine a resistance connected across the two terminals through a filter which passes only frequencies in a narrow band δf at f. If $G_R(f)$ is the power spectrum of the random voltage associated with the resistance, the mean rate of energy transfer from the resistance to the impedance will be

$$\frac{G_R(f)R(f)\,\delta f}{\{R+R(f)\}^2+X(f)^2}.$$

If $G_\rho(f)$ is the power spectrum of the random voltage associated with the impedance, the mean rate of energy transfer from the impedance to the resistance will be

$$\frac{G_\rho(f)R\,\delta f}{\{R+R(f)\}^2+X(f)^2}.$$

Since the resistance and impedance are supposed to be at the same temperature, the net mean rate of energy transfer must be zero or the second law of thermodynamics would be violated. Accordingly, equating the two rates of transfer, using the known expression for $G_R(f)$, and writing $\mathscr{R}[\rho(f)]$ for $R(f)$, we get

$$G_e(f) = 4\mathscr{R}[\rho(f)]kT. \qquad \ldots\ldots(86)$$

This constitutes an interesting and very useful extension of Nyquist's result.

4.5. *Recent Discussion of Nyquist's Results*

Fürth (1948) has given an interesting discussion of fluctuations associated with an impedance. He considers a network composed of an infinite number of inductances, capacities and resistances connected in a specified, quite simple, way across two terminals. The network is such that its impedance as a function of frequency may be made to have any desired form by choosing the elements suitably; it is also so chosen that the mean square value of the voltage fluctuations appearing across the terminals when nothing else is connected across them can be calculated by a straightforward statistical mechanical procedure of the type described earlier. Fürth shows that the mean square value of the voltage is related to the impedance by

$$\overline{V^2} = 4kT\int_0^\infty \mathscr{R}[\rho(f)]\,df. \qquad \ldots\ldots(87\,a)$$

Considering in a similar way an impedance formed from another network, he shows that if it is short-circuited the fluctuating current i through the impedance will satisfy

$$\overline{i^2} = \int_0^\infty \frac{4kT\mathscr{R}[\rho(f)]\,df}{|\rho(f)|^2}. \qquad \ldots\ldots(87\,b)$$

It may at first sight seem contrary to statements made in §4.2 that a statistical mechanical argument in which damping is ignored can arrive at results referring to the real part of the impedance. In fact, however, the values of the integrals in eqns. (87) depend only on the values of the inductances and capacities in Fürth's circuits and not at all on the values of the resistances in them. It should also be emphasized that the statistical mechanical argument does not justify the interpretation of the integrands in (87) as power spectra. Strictly speaking these results state nothing about the correlation function of the current or voltage fluctuations which might under suitable circumstances be determined from oscilloscope records. Thus the results (87), regarded as purely statistical mechanical, do not in our opinion contain the full content of Nyquist's result as obtained on the basis of the random force method and therefore do not obviate the need for the assumptions stated in §4.2: they do, however, establish that deductions from Nyquist's result agree with statistical mechanics in a wide range of circumstances.

Results analogous to Fürth's, referring to the fluctuations of any coordinate of a wide class of equilibrium systems, have been obtained by Callen and Welton (1951). They consider any system which has a range of densely spaced energy levels (an example would be the suspended system with its surroundings), choose a coordinate (e.g. the deflection defined microscopically) and obtain in a quite standard way an expression for the mean square fluctuation in that coordinate in terms of the matrix elements of that coordinate between the states

of the system. They then obtain in a very interesting way an expression for the impedance associated with that coordinate (i.e. the ratio of the amplitude of a sinusoidally varying generalized force to the amplitude of the resulting variation of the time derivative of the coordinate). They first find the average rate of transfer of energy to the system when it is initially in a definite quantum state and is perturbed by the sinusoidally varying generalized force, and then average this over an ensemble of initial states to get the actual rate of transfer of energy to the system. The impedance is at once derived from this rate of supplying energy. A somewhat more direct derivation of the expression for the impedance has been given by Jackson (1952). Callen and Welton then show in effect that the mean square value of the fluctuations in the coordinate Q and the impedance $\rho(f)$ are related by

$$\overline{\Delta Q^2} = \int_0^\infty \frac{4\mathscr{R}[\rho(f)]}{|\rho(f)|^2} [\tfrac{1}{2}hf + hf\{\exp(hf/kT)-1\}^{-1}]\, df \qquad \ldots\ldots(88)$$

and the classical limit of this corresponds to (87 *b*). Zero point fluctuations are included in (88). The result is purely statistical mechanical and here again an assumption is involved in interpreting the integrand as giving the power spectrum of the fluctuations of Q. Callen and Welton introduce the idea of a random force early in their argument and arrive at an expression for the mean square value of this rather than of ΔQ: they obtain the right-hand side of (88) with $|\rho(f)|^2$ as an additional factor in the integrand.

More recently Callen and Greene (1952) have used a very interesting argument to express the actual correlation function, and so the power spectrum, of the fluctuations of a coordinate of an equilibrium system in terms of the corresponding impedance. The crux of their argument is the relation discussed in the first section between the correlation function and the form of the macroscopic decay of the system. Callen and Greene, however, justify this relation by a microscopic argument which does not introduce the idea of the rapidly fluctuating force which we used in our simple discussion.

4.6. *Upper Limit to Frequencies for which Nyquist's Formula is Valid*

For very high frequencies for which $hf \sim kT$ it is, as Nyquist pointed out, necessary to use not the equipartition result but the quantum statistical modification of it. This leads to

$$G_R(f) = 4R\frac{hf}{\exp(hf/kT)-1} \qquad \ldots\ldots(89)$$

which indicates a falling off of the spectrum at very high frequencies ($\sim 10^{13}$ sec^{-1} at room temperature). This falling off is sufficient to make the integral giving the mean square voltage across the resistance finite. However, the physical considerations which led to assuming a short correlation time for the fluctuating voltage suggested that there would be an interval τ' over which correlation would be appreciable. This, by the Wiener–Khintchine relation (41*b*), implies a falling off in the spectrum at frequencies of order $1/\tau'$, as is borne out by a detailed kinetic theory analysis of voltage fluctuations carried out by Bakker and Heller (1939). This falling off is quite distinct from the falling off predicted by (89), although the frequencies at which both occur (the frequency limits) have, as it happens, the same order of magnitude.

The existence of the kinetic theory limit seems at first sight to be ruled out by the statistical mechanical argument given. It must be remembered however that at the high frequencies for which the question of a kinetic theory limit arises, the resistance will vary with frequency because of both ordinary capacity effects and effective capacities due to transit time, the latter depending on τ'. It is therefore possible that the falling off of the spectrum may simply correspond to a falling off of the real part of the impedance of the resistor. This is not ruled out by the Bakker and Heller argument, for they compared the spectrum with the spectrum corresponding to the low frequency resistance. Finally, one has to consider the fact that the notions of resistance and impedance may become inapplicable at very high frequencies. Most of the points made here concerning the upper limit have been made by Lawson and Uhlenbeck (1950, chap. 4).

4.7. *Systems acted on by a Number of Independent Fluctuating Forces*

Up to this point we have confined our attention in this section to systems acted on by only one random force. The basic result required in discussing the response to a number of independent random forces is that the power spectrum of the sum of two or more ***independent*** random functions is the sum of the power spectra of the random functions separately. It is convenient to establish a slightly more general result. Using an obvious notation, one gets from (41 *b*)

$$G_{av_1+bv_2}(f) = 4\int_0^\infty \overline{\{ay_1(t)+by_2(t)\}\{ay_1(t+\tau)+by_2(t+\tau)\}}\cos 2\pi f\tau\, d\tau$$

$$= a^2 G_{v_1}(f) + b^2 G_{v_2}(f). \qquad \ldots\ldots(90)$$

Accordingly if the system is linear, as will be assumed, we can determine the power spectrum of the fluctuations of the quantity of interest arising from each random force acting alone and then simply sum the power spectra. Alternatively we may work directly from the correlation functions of the random forces by working with the differential equations giving the response of the system to the various forces.

Some confusion has at times arisen as to the effect of the existence of various fluctuating forces on the magnitude of the fluctuations of the system. It is, however, important to note that if the system concerned is an equilibrium system, the magnitudes of the fluctuations are determined by statistical mechanical arguments, and there is then no question of augmenting them by adding fluctuating forces, provided the system remains an equilibrium system. From the random force point of view the reason is that introducing new sources of random forces inevitably introduces damping and the damping just cancels out the tendency of the new random force to augment the fluctuations. This has been pointed out very clearly by Zernike (1926). Confusion arose from the fact that one obtained the equipartition value for the mean square deflection of a galvanometer by considering either molecular bombardment alone or current fluctuations alone. But to obtain these results we must ignore the electromagnetic damping when we ignore the current fluctuations, and ignore the air damping when we ignore the molecular bombardment. The point to be borne in mind is the fixed relation between fluctuating forces and associated damping. Whenever we find a fluctuating force acting on an equilibrium system it must be possible to find an associated damping and vice versa.

As an example we may consider galvanometer fluctuations in a little more detail. If $P(t)$ and $E(t)$ are the random couple and e.m.f. respectively, the

equations governing the motion of the system are, with the neglect of inductance,

$$I\frac{d^2\theta}{dt^2}+\kappa_0\frac{d\theta}{dt}+c\theta=P(t)+Gi \qquad \ldots\ldots(91)$$

$$Ri=E(t)-G\frac{d\theta}{dt}$$

whence

$$I\frac{d^2\theta}{dt^2}+\left(\kappa_0+\frac{G^2}{R}\right)\frac{d\theta}{dt}+c\theta=P(t)+\frac{G}{R}E(t). \qquad \ldots\ldots(92)$$

The system therefore behaves as if the damping were κ_0+G^2/R and it were acted on by a couple of power spectrum given by

$$G_P(f)+\frac{G^2}{R^2}G_E(f)=4kT\left(\kappa_0+\frac{G^2}{R}\right). \qquad \ldots\ldots(93)$$

The ratio of the constant value of this power spectrum to the damping is fixed and this, from the random force point of view, is essentially the reason for the mean square fluctuation being independent of the amounts of mechanical and electromagnetic damping. It is a straightforward matter to determine from (92) the impedance function relating couple to deflection. The power spectrum of the fluctuations of θ can then be found from (42) and by integration we show that $\overline{\Delta\theta^2}$ has its equipartition value. This argument has been given by Van Lear (1933) and Surdin (1949). The direct correlation function approach can also be used to deal with (91): reference may be made to the paper by Jones and McCombie (1952), where the original argument given by Ornstein *et al.* (1927) is extended: a similar treatment is also given there which shows that the same result is obtained for $\overline{\Delta\theta^2}$ even if an inductance is included in the circuit of the galvanometer.

4.8. *Systems not in Equilibrium*

An excellent survey of the various sources of fluctuating forces or 'noise' in systems which are not in equilibrium (e.g. electronic devices of various sorts, semiconductors carrying currents, and so on) is included in the review article by MacDonald. We shall be content to indicate very briefly some of the more important sources of noise.

One of the simplest examples of a non-equilibrium system is an electrical network with resistances at differing temperatures. The temperature matters only at the resistances, since there is no interaction with the surroundings at inductances and capacitances in so far as these elements are ideal. Ornstein *et al.* (1927) included a cooled resistance in the circuit of a galvanometer and found the fluctuations of the galvanometer reduced below the equipartition value. The variation of amplitude with temperature conformed to the predictions of random force calculations, the random force associated with each resistance being of course that appropriate to the actual temperature of that resistance.

Of much greater practical importance are various electronic devices. We have already indicated (§3) the origin of 'shot noise'. This has a constant power spectrum up to a frequency of the order of the reciprocal of the transit time for the electron in the tube. Interaction of the electrons with the grids may induce noise which has not a constant spectrum, e.g. induced grid noise. Again the random sharing of electrons between the grids may give noise. In general these sources of noise will not be independent and in some cases suitable circuitry may help to cancel them out. As already mentioned, these problems are surveyed by

MacDonald. An extended account of them is given by Goldman (1948), Lawson and Uhlenbeck (1950), Schremp (1948), and an elementary discussion by 'Cathode Ray' (1952). For a recent detailed review see van der Ziel (1952).

In recent years a considerable amount of attention has been devoted to noise, greatly in excess of that predicted by Nyquist's formula, which arises in semiconductors carrying current. This noise resembles in some respects the excess noise above shot noise encountered in thermionic valves (flicker effect), and the relation between the origins of noise in these two situations has been extensively discussed. We shall give a brief review of some of this work.

Voltage fluctuations would be expected to occur in a semiconductor carrying a current because of fluctuations in resistance consequent upon fluctuations in the number of electrons in the conduction band. The appropriate analysis has been carried through by Bernamont (1937). Fluctuations in resistance will result in a mean square fluctuating voltage proportional to the square of the steady current flowing; this form of dependence on the current is fairly typical of the types of noise being considered here. If we assume that the carriers have a mean lifetime τ_1 in the conduction band, then an excess of carriers above the average number would on average decay to zero as $\exp(-t/\tau_1)$. By analogy with §2 it seems reasonable to assume that this will represent the form of the correlation function for positive t, so that

$$\overline{\Delta n(t)\,\Delta n(t+\tau)} \quad \propto \quad \exp(-|\tau|/\tau_1). \qquad \ldots\ldots(94)$$

It is not difficult to justify this result more fully by considering the average number of the carriers present at zero time which will be present at time t, and the average number which will have arrived during the interval 0 to t and remained till time t.

According to (41 *b*) the correlation function (94) implies that the power spectrum for the fluctuations will have the form

$$G(f) \propto \frac{\tau_1}{1+4\pi^2\tau_1^2 f^2}. \qquad \ldots\ldots(95)$$

Thus the power spectrum will be constant for low frequencies ($\ll 1/\tau_1$) and fall off as $1/f^2$ for higher frequencies. Essentially the same result has been obtained by Brillouin (1934) on considering the probability distribution of the carriers in phase space, and by Gisolf (1949) who considered the pulse of current caused by each carrier excited. The equivalence of these three approaches has been discussed by van der Ziel (1950). The effect is analogous to the shot effect. The larger the pulse of current given by each carrier while in the conduction band, the larger the noise. In the limit in which τ_1 is so large that each carrier which enters the conduction band traverses the semiconductor, the power spectrum of the noise will be just that corresponding to full shot noise. As follows from Campbell's theorem this has, for frequencies much less than the reciprocal of the transit time, a constant value depending only on the charge of the carrier e, and the magnitude of the current:

$$G(f) = 2eI. \qquad \ldots\ldots(96)$$

At low frequencies, however, the power spectrum of the observed noise greatly exceeds this full shot noise value. Consequently another origin must be sought for most of the observed current noise.

Part of the noise, above shot noise, observed in diodes is ascribed to fluctuations at the emitting surface (flicker effect). The similarity between flicker effect noise and excess noise in semiconductors suggests that part at

least of the latter noise may be due to effects at contacts similar to those at the emitting surface in a valve; it is likely that in some circumstances this contact noise will predominate. Schottky (1926) originally tried to explain the flicker effect by supposing that the rate of emission of electrons from the emitter surface depended on the concentration of ions on it and that the fluctuations in this number gave rise to the noise. However, the power spectrum for the fluctuations in the number of ions on the surface would have the same form as that for the fluctuations in the number of current carriers already considered. Thus one arrives at a $1/f^2$ law for all but the lowest frequencies, whereas experiment suggests a variation with frequency which is nearer $1/f$. The $1/f^2$ law will apply to any process which is characterized by a single time constant. It has been pointed out that the $1/f$ law can arise if the process is characterized by a range of time constants (Surdin 1951, van der Ziel 1950, du Pré 1950). These authors have shown that such a distribution of time constants is quite likely under a variety of circumstances. For example, the fundamental process might be the formation of clusters of impurity centres in the bulk material, or of ions at emitter surfaces and their subsequent disintegration by diffusion. The variety of configurations possible and the possibility of different ions or atoms being involved would give rise to a range of time constants. The application by Macfarlane (1947) of Sproull's (1945) theory of the build-up of cathode emission to explain flicker effect provides an example of the type of process under discussion, although it is complicated somewhat by the non-linearity of the mechanism involved. General comments on the possible processes are included in the papers by van der Ziel and Surdin.

Two detailed theories applicable to contact noise and flicker effect in diodes have been advanced more recently by Richardson (1950) and Macfarlane (1950). Macfarlane supposes that electron emission from a cathode takes place predominantly at a number of active patches on the surface and that the work function at such a patch depends strongly on the number of ions which have migrated along the surface on to the patch. He applies the theories of fluctuations in colloidal suspension, expounded by Chandrasekhar (1943), to derive the power spectrum of the fluctuations in the number of ions on a patch for various shapes of patch, and shows that a circular patch gives a power spectrum with the required frequency dependence. The theory will apply to contacts between grains, the areas of contact corresponding to the active spots.

Richardson supposes that the conductivity across a surface depends on the concentration of the ions at each point multiplied by some function of the position. By using a somewhat complicated form for the statistical properties of this weighting function he deduces the right frequency dependence for the fluctuations in conductivity due to movements of the ions on the surface. The mathematical apparatus set up in Richardson's paper is very powerful and general and should be useful in treating mathematically many linear processes giving rise to noise. The specific examples considered by Richardson are less simple and satisfactory than the model proposed by Macfarlane, but it is interesting to note that (as can easily be verified) Richardson's general methods when applied to Macfarlane's model lead to exactly the same result as the rather more specialized theory of colloid fluctuations used by Macfarlane.

Yet another source of current noise is the resistance fluctuations consequent on fluctuations in the temperature of the conductor. This effect is probably small in most work with semiconductors. It is discussed in §6.4.

§5. FEEDBACK

5.1. *Feedback with Suspended Systems*

We are now in a position to discuss the uncertainties in measurements with instruments incorporating feedback and the importance of the fact that such instruments do not constitute equilibrium systems. We supplement §2 with a discussion of the use of feedback in conjunction with suspended systems in order to make clear the type of circumstance under which the use of feedback may be advantageous, and, at the same time, to emphasize the existence of limits of accuracy which the use of feedback certainly will not enable us to transcend.

In order to apply feedback to a suspended system we arrange that the system is acted on always by a couple, which is some function of the deflection and its time derivatives. This is usually achieved by using a light spot and photocells to produce a voltage proportional to the deflection, obtaining the required function of the voltage and its derivatives by a suitable circuit, and then applying the couple by either electrostatic or electromagnetic interaction (Schmitt 1938). It will be assumed that the mechanism providing feedback introduces no new fluctuating forces.

We have seen that if an accurate measurement of a couple is to be made with a suspended system it is necessary that any damping of the nature of that we have so far considered (e.g. that due to viscosity or electrical resistance) be made small, as a limit to the accuracy of a measurement made in a given time is set by the existence of the fluctuating couple associated with such damping. This limit is absolute, and exists quite independently of whether or not feedback is used. We also saw, however, that we might have difficulty in achieving this limit of accuracy with a convenient measurement procedure, if $\kappa_{\min} S$ was much less than $I_{\min}$; the method described in §2.2.2 for obtaining high accuracy in this case did not quite attain the optimum limit and moreover involved the inconvenience and sensitivity to extraneous disturbance associated with working with a much underdamped suspension. Under these circumstances it may be possible to use feedback to great advantage. In the first place we can use it to adjust in effect the values of I and c. If we arrange that the couple applied to the suspension is $a\theta + b\, d^2\theta/dt^2$, with a and b constants, the equation of motion is

$$I\frac{d^2\theta}{dt^2} + \kappa\frac{d\theta}{dt} + c\theta = P(t) + a\theta + b\frac{d^2\theta}{dt^2}. \qquad \ldots\ldots(97)$$

The system will then respond to applied couples exactly like an equilibrium suspended system with constants $I-b$, κ and $c-a$. That the fluctuations of the system with feedback will be the same as those of this equilibrium system follows from the fact that the fluctuating couples in the two cases will be the same, namely that appropriate to the damping κ. The application of this procedure may be limited by stability considerations but it enables us in principle to work with a critically damped system while achieving the accuracy corresponding to the small value of κ, just as if we could actually reduce I. If we make b positive and $(I-b)/\kappa$ much less than the observation time, we can attain the optimum accuracy by the integration procedure of §2.2.3.

It is also possible to use differential feedback, which has the effect of changing the damping, since if the feedback couple is $-g\, d\theta/dt$ the damping constant in the equation of motion becomes $\kappa+g$. So far as the response to an applied couple

is concerned the system behaves as if its constants were I, $\kappa+g$ and c, but the fluctuating couple is still that appropriate to κ, not to $\kappa+g$. It follows that the magnitude of the power spectrum of the fluctuations in θ for every frequency, and so of the correlation function for every τ, will be reduced below that for an equilibrium system with constants I, $\kappa+g$ and c by a factor $\kappa/(\kappa+g)$. In particular $\frac{1}{2}c\,\overline{\Delta\theta^2}$ will be $\frac{1}{2}\boldsymbol{k}T\kappa/(\kappa+g)$. This introduction of damping without a corresponding fluctuating force, which has been commented on by Strutt and van der Ziel (1942), is possible only because the system with feedback is not in equilibrium: the possibility is ruled out for equilibrium systems by Nyquist's theorem and the analogues of it which apply to other damping forces.

Using differential feedback is usually a more convenient way of facilitating the attainment of the optimum accuracy than is the use of feedback to alter in effect the constants I and c. We can use differential feedback to make $(\kappa_{\min}+g)S$ as large compared with I as we please so as to enable us to use any of the methods of measurement discussed. If we had increased the damping to $\kappa+g$ by introducing air damping or electromagnetic damping, the accuracy attained would have been that appropriate to damping $\kappa+g$ but because of the reduced fluctuations associated with the differential feedback damping, the mean square error has to be multiplied by a factor $\kappa/(\kappa+g)$ and so we in fact attain the optimum accuracy appropriate to the damping κ.

MacDonald (1948) quotes van Zolingen (1948) as using differential feedback in conjunction with an electrometer. In this way he could attain in measuring a current the accuracy appropriate to the very small electromagnetic and air damping (the latter being reduced by evacuation), and yet retain the advantages of working with a critically damped system. If air damping were made negligible, the accuracy attained by the integration method would be $\overline{\Delta i^2}=2\boldsymbol{k}T/RS$, with R the very large leakage resistance of the instrument.

As a final example of the use of differential feedback we may consider the use of a galvanometer to measure a current in a circuit of resistance very much greater than the critical damping resistance of the galvanometer. Again one could get near the optimum accuracy by the method of §2.2.2, but working with an underdamped instrument in this way may not be convenient. By using differential feedback, the damping can be increased to critical without introducing other fluctuating forces, and the accuracy corresponding to the large circuit resistance attained. The same end could be achieved by using feedback to alter I and c, keeping their ratio fixed, but stability would probably be better assured with the differential feedback method.

5.2. *General Considerations on Feedback*

Let us now consider the use of feedback in conjunction with any measuring instrument. In so far as this is an ideal instrument, in the sense that no fluctuating forces have to be considered other than those introduced at the input, the use of feedback merely changes the impedance function (and possibly the input impedance) of the instrument. Suppose that the impedance function without feedback is $Z(f)$, and we arrange that the component of the output of frequency f produces a component at the input equal to this output component multiplied by $k(f)$, where $k(f)$ like $Z(f)$ is complex. Then, if a signal $A_{\mathrm{I}}(f)$ of frequency f produces an output $A_0(f)$, these are related by

$$A_0(f)=Z(f)\{A_{\mathrm{I}}(f)+k(f)\,A_0(f)\},$$

so that the new impedance function is given by

$$A_0(f) = \frac{Z(f)}{1 - k(f)Z(f)} A_1(f). \qquad \ldots\ldots(98)$$

As will be explained in § 7, changing the impedance function may be useful in diminishing the effect of the input fluctuating force, but it does not introduce novel features peculiar to feedback.

In general, however, measuring instruments such as electronic amplifiers have internal fluctuating forces. These give rise to an apparent increase in the fluctuating force appearing at the input. If $Z_1(f)$ is the impedance function between the output and the point where an internal fluctuating force of power spectrum $G_T(f)$ is appearing, and the only other fluctuating force is that with power spectum $G_I(f)$ at the input, the power spectrum $G_0(f)$ of the output fluctuation will be (the fluctuating forces being assumed independent)

$$G_0(f) = |Z(f)|^2 \left\{ G_I(f) + \frac{|Z_1(f)|^2}{|Z(f)|^2} G_T(f) \right\} \qquad \ldots\ldots(99)$$

and the expression in the curly brackets gives the apparent spectrum of the input fluctuating force. This fluctuating force is what sets a limit to the accuracy attainable, just as does a wholly real fluctuating force at the input. Now it has to be considered whether feedback might not alter this apparent power spectrum and so increase the attainable accuracy. With the feedback already discussed we get

$$G_0(f) = |Z(f)|^2 \{ G_I(f) + |k(f)|^2 G_0 f) \} + |Z_1(f)|^2 G_T(f),$$

i.e.
$$G_0(f) = \frac{|Z(f)|^2}{1 - |k(f)|^2 |Z(f)|^2} \left\{ G_I(f) + \frac{|Z_1(f)|^2}{|Z(f)|^2} G_T(f) \right\}. \qquad \ldots\ldots(100)$$

But the expression outside the brackets is the square of the modulus of the new impedance function, and the spectrum of the effective noise at the input is exactly as before. This important result is established by Strutt and van der Ziel (1942) who also give references to earlier applications of it. It applies only if the feedback is from the output back to the input, and does not apply to internal feedback. It is obvious, for example, that feedback increasing the amplification of stages prior to one at which an internal fluctuating force appears would decrease $|Z_1(f)|^2/|Z(f)|^2$ and so diminish the apparent input fluctuating force. For further details reference may be made to Twiss and Beers (1948) and to Strutt and van der Ziel (1942).

§ 6. RADIATION DETECTORS

A discussion of the uncertainties inherent in measurements with radiation detectors illustrates the use of power spectra and introduces a number of points of physical interest. Our main concern will be the methods of arriving at the theoretical sensitivity limits and no attempt will be made to discuss comprehensively how closely these limits have been approached by actual detectors.

6.1. *Fluctuations in Energy Transfer*

At least part of the uncertainty in a measurement with a radiation detector will arise from fluctuations in the rate of exchange of energy between the detecting element and the surroundings, and we must discuss the statistical properties of these fluctuations. The fluctuations in the rates of absorption and

loss of energy will be random functions denoted by $\phi_a(t)$ and $\phi_l(t)$ respectively. The difference $\phi_a(t)-\phi_l(t)$, the fluctuation in the net rate of gain of energy, is denoted by $\phi(t)$. We assume that these random functions have a negligible correlation time and that their power spectra are therefore constant. The constant value of G_ϕ, the power spectrum of ϕ, can be determined by a method analogous to that used in §3.5.1 where Nyquist's result was derived by considering a circuit with inductance and resistance. If we consider a body in equilibrium with its surroundings the fluctuations in temperature ΔT will satisfy

$$C\frac{d}{dt}(\Delta T)+\alpha\Delta T=\phi(t) \qquad \ldots\ldots(101)$$

where C is the heat capacity of the body and α the coefficient of heat transfer to the surroundings. We know from statistical mechanics (Schroedinger 1946) that the fluctuations in temperature must satisfy $C\overline{\Delta T^2}=kT^2$. On comparing these equations with $L\,di/dt+Ri=E(t)$ and $\overline{Li^2}=kT$ we obtain by analogy with Nyquist's result

$$G_\phi=4\alpha kT^2. \qquad \ldots\ldots(102)$$

To illustrate the significance of the power spectrum we may consider the case in which the body is a black body of area O. It then follows from Stefan's Law that α is $4O\sigma T^3$, and if we assume that the power spectra of ϕ_a and ϕ_l are each half of G_ϕ so that

$$G_{\phi_a}=G_{\phi_l}=2\alpha kT^2 \qquad \ldots\ldots(103)$$

it follows that G_{ϕ_a} is $8O\sigma T^5$. The mean square fluctuation in the energy absorbed in unit time will according to (63) be $\frac{1}{2}G_{\phi_a}$ or $4O\sigma T^5$. This reproduces a result obtained by Lewis (1947) from the statistics of a Bose–Einstein gas.

We shall now use (102) to consider the uncertainties in measurements with thermal detectors (thermocouples and bolometers) which operate by measuring the change in temperature produced in a receiving body by the radiation to be measured. An absolute limit can at once be obtained for the accuracy possible with thermal detectors in which (as in the thermocouple) the receiver is practically at the temperature of the surroundings. The best estimate of I_{abs}, the rate of absorption of energy, which can be made in a time S will include as an error the average of $\phi(t)$ over this time (cf. §2.1.1) and since this average has by (63) the mean square value $G_\phi/2S$ the minimum value for the mean square error in the estimate of the rate of *absorption* of radiation will be given by

$$\overline{\Delta I_{abs}^2}=2\alpha kT^2/S. \qquad \ldots\ldots(104)$$

If only a fraction ϵ_{ν_0} of the radiation to be measured (assumed to be monochromatic of frequency ν_0) is absorbed, the mean square error in the estimate of I_{inc}, the rate of incidence of the radiation, will, since $I_{inc}=I_{abs}/\epsilon_{\nu_0}$, be given by

$$\overline{\Delta I_{inc}^2}=2\alpha kT^2/S\epsilon_{\nu_0}^2. \qquad \ldots\ldots(105)$$

6.2. *The Thermocouple*

We shall now use (102) to calculate the power spectra of the fluctuations of current and voltage in a thermocouple circuit. It will be supposed that the thermocouple consists of two bodies (junctions) with heat capacities C_1 and C_2, and heat transfer coefficients α_1 and α_2, connected by wires (total resistance R_0) of different materials, such that if the two bodies are maintained at temperatures δT_1 and δT_2 above the surroundings there will be a thermal e.m.f. in the circuit

given by $P(\delta T_1 - \delta T_2)$, so that P is the thermoelectric power. It follows from the known relation between the Peltier coefficient and the thermoelectric power that if a current i flows round the circuit there will be, quite apart from Joule heating, heat generated at a rate $-PTi$ at the body 1, and at a rate $+PTi$ at the body 2: T is the absolute temperature of the circuit and its surroundings, δT_1 and δT_2 being assumed small. It will be assumed that the Joule heating is negligible in comparison, as is in fact the case for thermocouples in normal use (see Fellgett 1949 a).

If the direct transfer of heat from one junction to the other is not important the fluctuations in temperature and current will be governed by the equations

$$\left.\begin{aligned} C_1\frac{d}{dt}(\Delta T_1)+\alpha_1\Delta T_1=\phi_1-PTi, \quad C_2\frac{d}{dt}(\Delta T_2)+\alpha_2\Delta T_2=\phi_2+PTi \\ R_0 i=P(\Delta T_1-\Delta T_2)+E(t) \\ G_{\phi_1}(f)=4\alpha_1 kT^2,\ G_{\phi_2}(f)=4\alpha_2 kT^2,\ G_E(f)=4R_0 kT. \end{aligned}\right\} \quad .(106)$$

On replacing ϕ_1 in these equations by $\exp(2\pi ift)$, putting ϕ_2 and E zero and calculating the current i, the impedance function from input flux ϕ_1 to output current can be found, and the impedance functions from inputs ϕ_2 and E to output currents can be found similarly. The power spectrum $G_i(f)$ of the current fluctuations can then be found as the sum of the power spectra of the current fluctuations produced by ϕ_1, ϕ_2 and E separately (cf. § 4.7). The result can be expressed in the form

$$G_i(f)\,|R_0+\rho_D(f)|^2=4\mathscr{R}[R_0+\rho_D(f)]\,kT \qquad \ldots\ldots(107)$$

where

$$\rho_D(f)=P^2T\{\alpha_1+2\pi ifC_1\}^{-1}+P^2T\{\alpha_2+2\pi ifC_2\}^{-1}. \qquad \ldots\ldots(108)$$

Thus the fluctuations in current are just those which would be calculated from Nyquist's theorem if the circuit contained in addition to R_0 an impedance $\rho_D(f)$. This impedance $\rho_D(f)$ would in fact be included in a direct measurement of the impedance of the circuit: it is the 'dynamic impedance' arising from the fact that if a current flows in the circuit the temperature difference caused by Peltier heating and cooling at the junctions gives a back e.m.f. That the magnitude of this dynamic impedance is just $\rho_D(f)$ as given above can be shown by calculating the e.m.f. $P(\delta T_1 - \delta T_2)$ produced by a current $\exp(2\pi ift)$. Thus we see that the voltage fluctuations produced (via temperature variations) by fluctuations in the radiation exchange with the surroundings can be regarded as just the Nyquist voltage associated with a dynamic impedance. Once the dynamic impedance has been calculated the current fluctuations can be obtained by an application of Nyquist's result without any explicit reference to radiation fluctuations: the concept of dynamic impedance has been used in this way by Fellgett (1949 a) and Golay (1947, 1949): the latter treats his pneumatic detector in a similar way.

6.3. *Thermocouple as a Radiation Detector; Chopping of Radiation*

We shall suppose that the thermocouple is used as a radiation detector by including a current measuring device (of which the resistance will be included in R_0) in its circuit and measuring the current i produced by the radiation to be measured. This radiation will be supposed to be incident at rate I on junction 1 and to be completely absorbed there. The response time of the current measuring device will be assumed very long compared with the thermal response times C_1/α_1, C_2/α_2 of the thermocouple junctions so that it will respond appreciably only to frequencies much less than α_1/C_1, or α_2/C_2. For these frequencies the complex

part of ρ_D can be ignored : consequently ρ_D can be replaced by a resistance R_D given by

$$R_D = P^2T(1/\alpha_1 + 1/\alpha_2) \qquad \ldots\ldots(109)$$

and the Nyquist voltage associated with ρ_D can be replaced by that appropriate to R_D.

The relation between I and i can be obtained from the equations

$$\alpha_1\delta T_1 + iPT = I, \quad \alpha_2\delta T_2 - iPT = 0, \quad i = P(\delta T_1 - \delta T_2)/R_0 \qquad \ldots\ldots(110)$$

which lead to

$$I = i\alpha_1(R_0 + R_D)/P \text{ and } \overline{\Delta I^2} = \overline{\Delta i^2}\alpha_1^2(R_0 + R_D)^2/P^2. \qquad \ldots\ldots(111)$$

If the bandwidth of the current measuring instrument is B we can obtain $\overline{\Delta I^2}$ by replacing R by $R_0 + R_D$ in (67) and using (111) to get $\overline{\Delta I^2}$ from $\overline{\Delta i^2}$. This gives

$$\overline{\Delta I^2} = 4kTB\alpha_1^2(R_0 + R_D)/P^2. \qquad \ldots\ldots(112)$$

The smaller R_0, the smaller $\overline{\Delta I^2}$, and if R_0 could be made negligible we would have, using (109),

$$\overline{\Delta I^2} = 4kT^2B\alpha_1^2\left(\frac{1}{\alpha_1} + \frac{1}{\alpha_2}\right). \qquad \ldots\ldots(113)$$

It is desirable to have α_2, the heat transfer coefficient for the cold junction, large. This gives in the limit

$$\overline{\Delta I^2} = 4kT^2B\alpha_1. \qquad \ldots\ldots(114)$$

For an averaging process B is $1/2S$ and the optimum accuracy given by (105) is obtained. It is, however, sometimes necessary, in order to eliminate drifts, to make the two junctions identical and we then get $\overline{\Delta I^2} = 8kT^2B\alpha_1$.

If the measuring instrument is a galvanometer all the results of §2 can be used in a similar way. A particularly interesting case is that in which the galvanometer is underdamped and we block off the light on alternate half periods in order to use the resonance method of §2.2.3. The method is very useful in that the amplitude of the swings is very insensitive to slow 'drift' currents. It may be said that the signal to be measured has been labelled and the instrument so chosen that it amplifies preferentially incident radiation with this label.

Considerable attention has been devoted to comparing the errors introduced by fluctuations into measurements with the chopped light method with those introduced when the radiation is applied all the time. Here we shall compare the uncertainty in the resonance method, used with a galvanometer of negligible response time so that we can average the amplitude of the swings over the whole period, with the uncertainty in the integration method. Let us first suppose that α_2 is made large so that $R_D = P^2T/\alpha_1$. Combining (38) with (111) we get for the uncertainty with the resonance method

$$\overline{\Delta I^2} = \frac{\alpha_1^2}{P^2}\pi^2\frac{kT}{S}\left(R_0 + \frac{P^2T}{\alpha_1}\right). \qquad \ldots\ldots(115)$$

The uncertainty in an estimate in which the signal is applied all the time and the integration method of §2.2.3 used, is given by

$$\overline{\Delta I^2} = \frac{\alpha_1^2}{P^2}2\frac{kT}{S}\left(R_0 + \frac{P^2T}{\alpha_0}\right). \qquad \ldots\ldots(116)$$

This is about five times smaller than the limit with the resonance method, and gives the optimum result (105) if R_0 is negligible.

Suppose, however, that we use a symmetric thermocouple in which $\alpha_1=\alpha_2$. Then the dynamic impedance R_D is $2P^2T/\alpha_1$ and the mean square errors with both the resonance and integration methods are those just obtained multiplied by $(R_0+2P^2T/\alpha_1)/(R_0+P^2T/\alpha_1)$ i.e. doubled if R_0 is negligible. In using the resonance method, however, we need not simply cut off the light on alternate half-periods but can switch it from one junction to the other. Using (39) we obtain for this case

$$\overline{\Delta I^2}=\frac{\alpha_1^2}{P^2}\frac{\pi^2}{4}\frac{kT}{S}\left(R_0+\frac{2P^2T}{\alpha_1}\right). \qquad \ldots\ldots(117)$$

We see that if R_0 predominates (as is usually the case), the mean square error with the resonance method used in the way described is only $\pi^2/8$ times that with the integration method. If R_0 is negligible the mean square error with the resonance method is $\pi^2/4$ times the optimum mean square error (given by the integration method used with large α_2) and is $\pi^2/8$ times that obtained on applying the integration method to the symmetric thermocouple.

In order to obtain an idea of the order of magnitude of the uncertainty implied by (114) we may consider the junction to be a black body of area 0·4 mm², in radiative exchange with its surroundings. Then α_1 is given by $4O\sigma T^3$ and with a bandwidth of 1 c/s we get $\{\overline{(\Delta I)^2}\}^{1/2}=4\times10^{-12}$ watt. The best thermocouple of this area cited by Fellgett (1949 a) gave 3×10^{-11} watt for $B=1\,\mathrm{sec}^{-1}$. The difference will, in part at least, be because α_1 is increased by conduction along the wires and because R_0 is appreciable. The requirement that the conduction along the wires should be small conflicts with the requirement that R_0 should be small. Fellgett quotes expressions obtained by Daly (1947) for the best results attainable with given materials.

6.4. *The Bolometer*

The bolometer, unlike the thermocouple, is not an equilibrium system: the strip of which the change in temperature is to be measured carries a steady current in order to convert changes in resistance into changes in voltage. There is, therefore, no reason to expect that the fluctuations can be discussed in terms of Nyquist's result, as was possible with the thermocouple. The Nyquist voltage associated with the resistance of the strip and the fluctuating voltage consequent on fluctuations in the temperature of the strip have to be considered separately. The treatment given here is essentially that of Milatz and van der Velden (1943).

Let us suppose that the bolometer strip is a black body of area O, resistance r and temperature coefficient of resistance η, and that it carries a fixed steady current i. Energy incident at rate I will produce a change v in the voltage across the strip, given by

$$v=(I/4O\sigma T_b^3)i\eta r, \qquad \ldots\ldots(118)$$

if T_b is the temperature of the strip. The fluctuating voltage associated with the resistance r will have a power spectrum $4rkT_b$. Because of the heating by the current, T_b will exceed the temperature of the surroundings T_0, and this complicates the calculation of the fluctuations in the rate of exchange of energy with the surroundings. We can, however, assume that the power spectrum of ϕ_2, the fluctuations in the rate of absorption of energy, will be the same as if everything were at temperature T_0, while that of ϕ_1 will be the same as if everything were at temperature T_b: thus according to (103)

$$G_{\varphi_2}=8O\sigma kT_0^5,\ G_{\varphi_1}=8O\sigma kT_b^5 \text{ and } G_{\varphi}=8O\sigma k(T_0^5+T_b^5). \qquad \ldots\ldots(119)$$

The impedance function relating input ϕ to output temperature fluctuations can be obtained from (101) with α replaced by $4O\sigma T_b^3$: $G_{\Delta T}(f)$ can then be found and the power spectrum of the resulting fluctuations in v will be this multiplied by $i^2\eta^2r^2$. Adding the power spectrum of these fluctuations to the power spectrum of the Nyquist fluctuations gives

$$G_v(f)=\{4r\mathbf{k}T_b+8O\sigma\mathbf{k}(T_0^5+T_b^5)i^2\eta^2r^2[(4O\sigma T_b^3)^2+4\pi^2f^2C^2]^{-1}\}. \quad \ldots\ldots(120)$$

If the voltage measuring device has a narrow bandwidth B (much less than $4O\sigma T_b^3/C$), Δv^2 will be given by $G_v(0)B$. Expressing $\overline{\Delta I^2}$ in terms of Δv^2 by means of (118) we therefore get

$$\overline{\Delta I^2}=[4r\mathbf{k}T_b(4O\sigma T_b^3)^2/i^2\eta^2r^2+8O\sigma\mathbf{k}(T_0^5+T_b^5)]B. \quad \ldots\ldots(121)$$

For the averaging process considered by Milatz and van der Velden B is $1/2S$ and this reproduces their result.

If T_b were independent of i it would always be desirable to increase i and so make the Nyquist fluctuations have negligible effect. Increasing i, however, increases T_b since

$$O\sigma(T_b^4-T_0^4)=i^2r \quad \ldots\ldots(122)$$

and so will increase the second term: the first term has a minimum at $T_b/T_0\simeq 1{\cdot}24$. Following Milatz and van der Velden we can find the value of i which makes $\overline{\Delta I^2}$ a minimum. It turns out that if as is usually the case η is approximately $1/T_0$ (T_0 being about 300°K) the current should be adjusted till T_b/T_0 is about 1·2: making T_b/T_0 1·7 or 1·04 doubles $\overline{\Delta I^2}$.

Milatz and van der Velden point out that (as we have already seen) the reason the fluctuating voltage associated with the resistance cannot be made negligible by increasing the sensitivity to radiation is that sensitivity cannot be increased without increasing the temperature, and they suggest that a dielectric bolometer in which there would be little heating might overcome this difficulty. Considerations similar to those given in this section can be applied to bolometers in which there is conduction cooling (e.g. the evaporated metal bolometers described by Billings, Hyde and Barr (1947)). It may however be necessary where the bolometer is non-metallic (e.g. the Bell Telephone Laboratory semiconducting bolometer) to consider the possibility of the existence of current noise other than that due to temperature fluctuations. Usually the bolometer strip is incorporated in a bridge circuit along with an exactly similar compensating strip and there is increased noise associated with the presence of the second strip. The resonance method in which the light is applied to alternate strips for alternate half-periods is then almost as favourable as the integration method applied to the same system (cf. § 6.3).

6.5. *Limitations on Radiation Detectors in General*

We shall now consider the limits set by the fluctuating exchange of energy with the surroundings to the accuracy of measurement with an arbitrary radiation detector. We shall suppose that a fraction ϵ_ν of photons with frequency ν incident on the detector are effectively absorbed (e.g. excite an electron in a photoconductive cell) and that fluctuations in the emitted radiation are not detected: the extensions of the theory necessary if this last condition is not satisfied (as in a thermal detector not at zero temperature) will be obvious. We shall require to know the power spectrum of the fluctuations in the rate of

effective absorption of energy in the form of photons with frequencies in the range ν to $\nu+\delta\nu$. (The frequency ν of the electromagnetic vibration must be distinguished from the frequency f of a component in a Fourier analysis of the time variation of the rate of incidence of energy.) This power spectrum can be determined by a method due to Fellgett (1949 b). Consider a body which has absorptivity ϵ_ν for frequency in the range ν to $\nu+\delta\nu$ and zero absorptivity for other frequencies. The body is assumed in equilibrium with its surroundings which are at temperature T. The average rate of absorption or emission of photons by such a body will according to Planck's law be $\overline{m(\nu)}$, given by

$$\overline{m(\nu)}=\frac{\epsilon_\nu}{\mathbf{c}^2}\frac{2\pi O\nu^2\delta\nu}{\exp(\boldsymbol{h}\nu/\boldsymbol{k}T)-1} \qquad \ldots\ldots(123)$$

and α for the body will be given by $d(\boldsymbol{h}\nu\overline{m(\nu)})/dT$. On substituting in (103) we find that the power spectrum can be written $O\epsilon_\nu g(\nu, T)\delta\nu$ where $g(\nu, T)$ is a function of ν and T given by

$$g(\nu, T)=\frac{\boldsymbol{h}^2\nu^2 4\pi\nu^2\exp(\boldsymbol{h}\nu/\boldsymbol{k}T)}{\mathbf{c}^2\{\exp(\boldsymbol{h}\nu/\boldsymbol{k}T)-1\}^2}. \qquad \ldots\ldots(124)$$

It is assumed that $O\epsilon_\nu g(\nu, T)\delta\nu$ gives the power spectrum of the fluctuations in the rate of effective absorption of radiation energy in the range ν to $\nu+\delta\nu$ for *any* body with area O and effective absorptivity ϵ_ν exposed to radiation from surroundings at temperature T. According to (63) this is equivalent to the physically more perspicuous result that the mean square fluctuation in the number of photons in the range ν to $\nu+\delta\nu$ absorbed in unit time is

$$\overline{m(\nu)}\{1+[\exp(\boldsymbol{h}\nu/\boldsymbol{k}T)-1]^{-1}\}. \qquad \ldots\ldots(125)$$

This is to be compared with the mean square fluctuation $\overline{m(\nu)}$ to be expected if the photons could be regarded as a stream of independent particles, in which case the fluctuations would be Poisson fluctuations. If ϵ_ν were unity (125) would again reproduce a result obtained by Lewis (1947) but it is not clear how the direct statistical approach of Lewis would apply for $\epsilon_\nu\neq 1$. One might expect for example that the 'random' selection of photons for absorption would tend to make the fluctuations approximate more closely to Poisson fluctuations, but this is not so.*

Clark Jones (1947) showed that the similarities in the sensitivity limits of various detectors arise from the fact that these limits all originate in fluctuations in the exchange of radiation with the surroundings. Fellgett (1949 b) has pointed out, however, that the precise value of the resulting sensitivity limit depends on the nature of the detector. Let us consider first a *quantum detector* which responds in the same way to all *photons* effectively absorbed, independently of what their frequency may be. We suppose that the effective absorptivity of the detector is ϵ_ν and that it is being used to measure a steady stream of monochromatic radiation of frequency ν_0. Fluctuations of power spectrum $O\epsilon_\nu g(\nu, T)\delta\nu$ in the rate of absorption of radiative energy in the range ν to $\nu+\delta\nu$ from the surroundings will produce the same effect as a fluctuating flux of power spectrum $(\boldsymbol{h}^2\nu_0^2/\boldsymbol{h}^2\nu^2)O\epsilon_\nu g(\nu, T)\delta\nu$ of radiative energy of frequency ν_0. Thus the fluctuations in the rate of absorption of radiation from the surroundings will set the same limit to the accuracy of the estimate of the quantity being measured (which

* We are indebted to Dr. R. A. Smith for drawing our attention to the difficulties arising in the direct statistical approach when $\epsilon_\nu\neq 1$.

for the moment we take to be the rate of *absorption* of radiation) as would the existence of an indistinguishable fluctuating quantity of power spectrum

$$\int_0^\infty O\epsilon_\nu(\nu_0^2/\nu^2)g(\nu, T)\,d\nu.$$

If the whole detecting system, from input radiation to the observed output, has bandwidth B the mean square error in the estimate of the rate of *absorption* of radiation will (cf. § 3.7) be

$$B\int_0^\infty O\epsilon_\nu(\nu_0^2/\nu^2)g(\nu, T)\,d\nu.$$

The mean square error in the estimate of the rate of *incidence* of radiation will be $1/\epsilon_{\nu_0}^2$ times as great, i.e.

$$\overline{\Delta I^2} = \frac{BO}{\epsilon_{\nu_0}^2}\int_0^\infty \epsilon_\nu(\nu_0^2/\nu^2)g(\nu, T)\,d\nu \qquad \ldots\ldots(126)$$

where $g(\nu, T)$ is as defined in (124).

The other type of detector to be considered is an energy detector which responds in the same way to all *energy* effectively absorbed, independently of its frequency. In a discussion analogous to the above the factor ν_0^2/ν^2 will not appear and we would clearly obtain for the energy detector

$$\overline{\Delta I^2} = \frac{BO}{\epsilon_{\nu_0}^2}\int_0^\infty \epsilon_\nu g(\nu, T)\,d\nu. \qquad \ldots\ldots(127)$$

It is sometimes advantageous to chop the radiation and detect the component with the chopping frequency f_0. In this case the equations (126) and (127) are still significant and (cf. § 3·7) are to be interpreted as giving the mean square values of the components with frequency f_0 which will produce an output with the same mean square value as the noise output. They thus give in effect what is conventionally termed the minimum detectable signal.

In applying the results embodied in (126) and (127) to actual detectors, it is necessary to know (*a*) whether the detector is a quantum or an energy detector, (*b*) whether it records fluctuations in the emitted as well as the absorbed radiation, (*c*) the area O, and (*d*) the effective absorptivity ϵ_ν. We assume that (*a*) and (*b*) are known from the theory of the mechanism involved, and that (*c*) can be obtained experimentally. The effective absorptivity presents difficulties. A function of ν to which it is proportional can be found by observing experimentally the spectral sensitivity A_ν defined by $V_\nu = A_\nu I_\nu$, where V_ν is the output produced by a rate of incidence I_ν of energy of frequency ν. Thus, for an energy detector $V_\nu = a\epsilon_\nu I_\nu$ (where a is a frequency independent constant) so that $\epsilon_\nu = A_\nu/a$: for a quantum detector $V_\nu = b\epsilon_\nu I_\nu/h\nu$ so that $\epsilon_\nu = h\nu A_\nu/b$. The estimates given for $\overline{\Delta I^2}$ in (126) and (127) increase with a and b respectively. If we now fix the proportionality constant in the appropriate expression for ϵ_ν by choosing a or b to be as small as possible, the limit being set by the fact that the maximum value of ϵ_ν must not exceed unity, we shall find, on using in (126) or (127) the value of ϵ_ν so obtained, a lower limit for the uncertainty with a detector of the assumed mechanism with the observed area O, and spectral response A_ν.

Examples of the application of (126) and (127) to specific instruments will be found in the paper by Fellgett (1949 b). Further examples are considered by Moss (1950). He finds that for lead telluride photoconductors the uncertainty due to fluctuations in the incident radiation increases as the cell is cooled because of the

change in spectral response. However Müser (1951) emphasizes that this does not imply that there will actually be an increase in the observed noise on cooling, since there will be noise due to fluctuation in the excitation of electrons by lattice vibrations and this will diminish on cooling.

The quoted theoretical limiting values (obtained in the way described from the experimental spectral response curve) for $(\overline{\Delta I_{\nu_0}{}^2})^{1/2}$ for photoconductive cells (lead sulphide or lead telluride) of areas 0·1 cm² are of the order of 10^{-13} watt (B being taken equal to 1 sec^{-1} and it being assumed that ϵ_{ν_0} is unity). Actual cells cooled to 90°K but exposed to room temperature radiation get to within a factor of two or three of this limit, and some may do much better. The values of $(\overline{\Delta I_{\nu_0}{}^2})^{1/2}$ are ten or a hundred times less than for a black body detector of the same area. The limits can be improved by shielding the photocell, with cooled shields, from as much of the room temperature radiation as possible.

As emphasized by Fellgett, the photoconductor excels the black body thermal detector only because its frequency response is more selective so that only the components in the fluctuating radiative exchange with values of ν lying in a restricted range contribute to the noise. A radio receiver attains a very much smaller value for $(\overline{\Delta I_{\nu_0}{}^2})^{1/2}$ since the values of ν which are effective are restricted by the selectivity of the receiver. This point of view is ratified by using it to reproduce the result obtained from Nyquist's theorem that the energy transferred through a filter of bandwidth B from an aerial of radiative resistance R to a matched load is $\boldsymbol{k}TB$. It is known (see Fellgett 1949 b) that the energy in the range B at ν absorbed by the aerial and transmitted to a matched resistance is the same as that absorbed by a black body of area $\boldsymbol{c}^2/2\pi\nu^2$. Using the expression (123) for the mean rate of absorption of photons by such a body we get for Φ_a, the mean rate of absorption of energy by the aerial,

$$\overline{\Phi}_a = \frac{\boldsymbol{h}\nu 2\pi\nu^2}{\boldsymbol{c}^2\{\exp(\boldsymbol{h}\nu/\boldsymbol{k}T)-1\}}\frac{\boldsymbol{c}^2}{2\pi\nu^2}B = \frac{\boldsymbol{h}\nu}{\exp(\boldsymbol{h}\nu/\boldsymbol{k}T)-1}B. \quad \ldots\ldots(128)$$

This is just $\boldsymbol{k}TB$ if $\boldsymbol{h}\nu \ll \boldsymbol{k}T$: in fact (128) corresponds to the more exact quantum mechanical form of Nyquist's theorem. The fluctuations in Φ_a have a constant power spectrum $(\boldsymbol{c}^2/2\pi\nu^2)g(\nu, T)B$ (cf. eqn. (124)). Assuming that B gives the range of f to which the circuit responds, the minimum detectable signal in the sense of §3.7 will be given in the classical limit by

$$\overline{(\Delta\Phi_a)^2} = 2\boldsymbol{k}^2T^2B^2. \quad \ldots\ldots(129)$$

This result can also be obtained from Nyquist's result together with the assumption that the distribution of values of the fluctuating voltage is gaussian, so that $\overline{V^4} = 3(\overline{V^2})^2$. For B equal to 1 sec^{-1} eqn. (129) gives about 6×10^{-21} watt for $\{(\Delta\Phi_a)^2\}^{1/2}$, and this gives the order of the minimum detectable rate of absorption of energy.

§7. DISCRIMINATION AGAINST NOISE

7.1. *Wiener's Criterion*

Let us suppose that a quantity (the signal), of which the variations are of interest, is accompanied by a fluctuating quantity of the same kind (noise). The signal and noise will appear together at the input of any instrument used for following the signal. If the reproduction is to be good the instrument must have a response time short enough to allow it to follow the details of the signal. On the other hand a short response time implies an impedance function which

is large over a wide range of frequencies and if, as if often the case, the noise has a practically constant power spectrum, a broad passing band will mean that a large amount of the noise will be amplified. Thus compromise is necessary.

In order to formulate the problem more precisely, let us suppose that we should like the output $o(t)$ and the signal $s(t)$ to satisfy the relation

$$o(t) = As(t) \qquad \ldots\ldots(130)$$

where A is some *prescribed* constant (the amplification). The presence of noise makes it impossible to satisfy this relation exactly and the difference $o(t) - As(t)$ may be termed the error. Wiener's (1949) criterion for the choice of the impedance function of the instrument used to follow the signal is that it should be such as to make the mean square value of this error as small as possible. Other criteria are possible and might in particular circumstances be more appropriate, but Wiener's criterion is very useful and is particularly amenable to analysis.

In the present context we regard the signal as a random function and, in particular, ascribe to it a power spectrum. The appropriate form for the power spectrum will be determined by the type of signal the instrument is to be designed to follow. In order to determine the power spectrum experimentally we would apply samples of the signal to be measured (together with the noise) to the input of an instrument with a passing band including all frequencies for which the power spectrum of the signal is found to be large. The output would be recorded and the power spectrum of the output could be estimated from the resulting records (cf. §3.3). Just what will constitute a sample of the signal sufficient to give a typical length of output will be clear from the homogeneity of the records obtained. Thus, for example, the records of the outputs obtained from a considerable range of compounds might be needed to determine a suitable amplifier to associate with an infra-red spectrometer. Dividing the estimated power spectrum of the output by the square of the modulus of the impedance function gives an estimate of the power spectrum of signal plus noise. A similar analysis of records obtained in the absence of signal will lead to an estimate of the power spectrum of the input noise and, the signal and noise being assumed independent, subtraction will give an estimated power spectrum of the signal.

We shall now obtain an expression for the impedance function satisfying Wiener's criterion in terms of the power spectra of signal and noise. From this expression we can find the minimum mean square value of the ratio $\overline{\{o(t) - As(t)\}^2}/A^2\overline{s^2(t)}$ for the given power spectra of signal and noise: the value of this will indicate whether it is worth trying to measure the signal.

As in §3.3 we introduce a very long period t_0 and use as random variables the coefficients of Fourier analyses of the signal $s(t)$ and noise $n(t)$ so that, measuring the signal from its mean value,

$$s(t) = \sum_{r=1}^{\infty} \{a_{s,r} \cos(2\pi rt/t_0) + b_{s,r} \sin(2\pi rt/t_0)\}, \qquad \ldots\ldots(131)$$

$$n(t) = \sum_{r=1}^{\infty} \{a_{n,r} \cos(2\pi rt/t_0) + b_{n,r} \sin(2\pi rt/t_0)\}. \qquad \ldots\ldots(132)$$

We suppose that the impedance function is $Z(f_r) = x_r + iy_r$. Then the output $o(t)$ will be given by

$$o(t) = \sum_{r=1}^{\infty} [\{(a_{s,r} + a_{n,r})x_r + (b_{s,r} + b_{n,r})y_r\} \cos(2\pi rt/t_0) + \{(b_{s,r} + b_{n,r})x_r - (a_{s,r} + a_{n,r})y_r\} \sin(2\pi rt/t_0)]. \qquad \ldots\ldots(133)$$

The error is $o(t)-As(t)$. Using (43) with $\tau=0$, and remembering that the random variables corresponding to signal and noise are independent of each other, we get

$$\overline{\{o(t)-As(t)\}^2}=\tfrac{1}{2}\sum_{r=1}^{\infty}[(\overline{a_{s,r}^2+b_{s,r}^2}+\overline{a_{n,r}^2+b_{n,r}^2})(x_r^2+y_r^2)-2A(\overline{a_{s,r}^2+b_{s,r}^2})x_r + A^2(\overline{a_{s,r}^2+b_{s,r}^2})]. \qquad \ldots\ldots(134)$$

The minimization is particularly simple and letting t_0 tend to infinity and writing $G_n(f_r)$ for $\frac{1}{2}t_0(\overline{a_{n,r}^2+b_{n,r}^2})$ we get for the optimum response function $Z_0(f_r)$,

$$x_r=\frac{AG_s(f_r)}{G_s(f_r)+G_n(f_r)},\ y_r=0. \qquad \ldots\ldots(135)$$

Substituting in (134) and replacing summation by integration we get for the minimum fractional error

$$\left[\frac{\overline{\{o(t)-As(t)\}^2}}{A^2\overline{s^2(t)}}\right]_{\min}=\frac{\int_0^\infty \frac{G_s(f)G_n(f)}{G_s(f)+G_n(f)}df}{\int_0^\infty G_s(f)df}. \qquad \ldots\ldots(136)$$

This result has been obtained by Wiener (see below).

We have placed no restrictions on the impedance function and it will not be possible to construct an instrument with the impedance function (135): to determine the output at time t corresponding to this impedance function one requires to know the form of the input both before and after t. Given the output of an arbitrary instrument, however, of known $Z(f)$ and passing band at least of the order of that given by (135), we can, by multiplying a Fourier series analysis of the record by $Z_0(f)/Z(f)$, determine the output which would have been obtained with the impedance function (135); in practice it would be more convenient to find an equivalent procedure whereby the knowledge of the record round each point will enable us to obtain the optimum output at that point.

The use of the impedance function $Z_0(f)$ in this way is not very practicable and is quite useless if an instantaneous estimate of the signal is required—as when the signal is operating a servo-mechanism. Wiener (1949) has solved the problem of determining the optimum response under the condition that we know the input only up to a time $t+\tau$ where t is the instant for which the estimate of the value is required. If τ is positive this corresponds to working with a time lag, if τ is negative it implies working with a predictor. We shall not attempt to discuss Wiener's analysis here. Expositions of it, and procedures for approximating to the optimum response function have been discussed by Levinson in appendices to Wiener's book. Our result agrees with the one obtained by Wiener in the limit $\tau\to\infty$ as it clearly should.

7.2. *Optimum Response of Critically Damped Galvanometer*

In order to get some idea of the significance of the results obtained from the Wiener criterion we have used it to determine the disposable constants (zero frequency amplification and response time) of a critically damped galvanometer used to follow a signal in the presence of noise where the power spectra $G_s(f)$ and $G_n(f)$ of signal and noise respectively are

$$G_s(f)=\frac{4b}{b^2+4\pi^2f^2},\ G_n(f)=2c, \qquad \ldots\ldots(137)$$

c and b being constants. The power spectrum $G_s(f)$ is that corresponding to the simple correlation function $\exp(-|\tau|/\tau')$, with the lifetime τ' equal to $1/b$. Three values of c were chosen which made the optimum values for the ratio mean square error to mean square signal given by (136), about 1, 10 and 40%. The calculation was effected by writing (134) in integral form, and substituting for $x+iy$ the impedance function of the galvanometer expressed in terms of the response time τ_0 and the zero frequency amplification k_g; the expression could be minimized with respect to k_g algebraically and the minimum of the resulting expression with respect to τ_0 was found graphically. We are indebted to Mr. G. A. Ingram who carried through these calculations. The results are summarized in the table.

The values of τ_0' and τ_0'' did not exist with the largest error ($\mathscr{E}_0 = 0{\cdot}643$) since the fractional error is made equal to unity by dropping k_g to zero so that $o(t)$ becomes zero. We include also for comparison the optimum value $\mathscr{E}_w$ for a realizable instrument with zero time lag as obtained from Wiener's results and the absolute optimum $\mathscr{E}_0$ according to (136).

c	k_g/A	τ_0/τ'	τ_0'/τ'	τ_0''/τ'	$\mathscr{E}_0$	$\mathscr{E}_w$	$\mathscr{E}_0$
$1/5b$	0·357	1·05	—	—	0·64	0·58	0·41
$1/100b$	0·787	0·31	1·57	0·055	0·22	0·18	0·10
$10^{-4}/b$	0·976	0·031	0·126	0·0055	0·022	0·020	0·01

$\mathscr{E}_0$ = minimum value of fractional mean square error using a galvanometer, τ_0/τ' = ratio of optimum response time to lifetime τ', k_g/A = ratio of amplification at zero frequency to amplification A, τ_0' and τ_0'' are the values of τ_0 for which the error is twice the minimum.

It will be seen that the choice of response time is not very critical but that departure from the optimum by a factor of about five doubles the fractional mean square error. The result obtained with a critically damped galvanometer is not much inferior to Wiener's and for low noise both are about twice the absolute optimum value. These results are rather limited but they probably give a good idea of the sort of importance which should be attached to choice of bandwidth or impedance function. It should perhaps be pointed out that if there is noise in a localized range of frequencies it may be worth going to considerable trouble to get the impedance function small in that range.

7.3. *Optimum Response to a Pulse*

Another problem is the choice of an instrument to measure the magnitude of a pulse of known shape in the presence of noise of power spectrum $G_n(f)$. In this case we shall require that the ratio of the peak of the response to the pulse to the root mean square noise output should be as large as possible. The problem has been treated by den Hartog and Muller (1947), and references to American workers are given by Uhlenbeck and Lawson. We can again use the Fourier series method. In order that there may be nothing peculiar about the point $t=0$ let us choose the interval from $-\frac{1}{2}t_0$ to $+\frac{1}{2}t_0$. Let there be an input pulse of the given form $P(t)$ in each period and let it be represented by

$$P(t) = \tfrac{1}{2}a_0 + \sum_{r=1}^{\infty} (a_r \cos 2\pi rt/t_0 + b_r \sin 2\pi rt/t_0). \qquad \ldots\ldots(138)$$

If the response function $Z(f_r)$ is written x_r+iy_r then the value of the output deflection $P_0(t)$ at the zero of time will be

$$P_0(0) = \tfrac{1}{2}a_0x_0 + \sum_{r=1}^{\infty} (a_rx_r + b_ry_r). \qquad \ldots\ldots(139)$$

The noise output will be (using (63) to get the first term)

$$\overline{n_0^2} = \frac{G_n(0)x_0^2}{2t_0} + \frac{1}{t_0}\sum_{r=1}^{\infty} G_n(f_r)(x_r^2+y_r^2). \qquad \ldots\ldots(140)$$

The values of x_r and y_r corresponding to the optimum response function can be obtained by requiring that $P_0(0)$ should be a maximum when $\overline{n_0^2}$ has some fixed value. Using the method of undetermined multipliers this gives, with λ a constant determined by the prescribed (arbitrary) noise level,

$$x_r = \lambda t_0 \frac{a_r}{G_n(f_r)}, \qquad y_r = \lambda t_0 \frac{b_r}{G_n(f_r)}, \qquad r=0, 1, 2, \ldots \quad \ldots\ldots(141)$$

and this is the required result. In the particular case in which $G_n(f_r)$ is a constant, if an impulsive signal is applied to the optimum amplifier at time $t=0$, the output will be

$$\text{const.} \sum_{r=0}^{\infty} (a_r \cos 2\pi rt - b_r \sin 2\pi rt). \qquad \ldots\ldots(142)$$

This is just const. $P(-t)$ in agreement with the result given by den Hartog and Muller. In most applications it will be permissible to introduce sufficient lag into the detector to make possible a good approximation to the optimum response.

7.4. *Information Theory*

Our discussion of the operation of instruments attempting to follow or detect a signal in the presence of noise has been based on particular criteria of desirable performance. The more general statement of the function of the ideal instrument is that it should give as much information as possible about the signal as can be derived from the signal plus noise appearing at its input. The concepts of information theory (see Shannon 1948), enable one to give a quantitative significance to this requirement, which has recently been formulated very lucidly by Woodward and Davies (1952). They suppose that before the message is received we can assign an *a priori* probability distribution to the possible signals. After the signal (corrupted by noise) is received the additional knowledge will imply a different (*a posteriori*) probability distribution which will tend to be large for signals similar to the one actually received. They point out that the maximum amount of information we can hope to get from the instrument is this *a posteriori* probability distribution. The ideal instrument would calculate the probability distribution for us. Thus, for example, suppose we know that a single weak pulse appears somewhere in the course of a record, the *a priori* probability of appearance being the same for all points, the ideal instrument on receiving the pulse corrupted by noise would give us a plot of the *a posteriori* probability of the pulse being at various points.

ACKNOWLEDGMENTS

I am deeply indebted to Professor R. V. Jones for introducing me to the subject matter of this review and for his continued advice and encouragement. Special thanks are due to Mr. G. A. Ingram for his criticisms and help in preparing the manuscript. I am grateful to Mr. W. C. Jolly, Dr. H. P. Myers and Dr. W. Paul for their assistance in various ways, and to other members of the staff of the Natural Philosophy Department of Aberdeen University for discussions on various points.

REFERENCES

BAKKER, C. J., and HELLER, G., 1939, *Physica*, **6**, 262.

BARNES, R. B., and MATOSSI, F., 1932, *Z. Phys.*, **76**, 24.

BARNES, R. B., and SILVERMAN, S., 1934, *Rev. Mod. Phys.*, **6**, 162.

BERNAMONT, J., 1937, *Ann. Phys., Paris*, **7**, 71.

BILLINGS, B. H., HYDE, W. L., and BARR, E. E., 1947, *J. Opt. Soc. Amer.*, **37**, 123.

BRILLOUIN, L., 1934, *Helv. Phys. Acta*, **7**, 47 (Suppl.).

CALLEN, H. B., and GREENE, R. F., 1952, *Phys. Rev.*, **86**, 702.

CALLEN, H. B., and WELTON, T. A., 1951, *Phys. Rev.*, **83**, 34.

CATHODE RAY, 1952, *Wireless World*, **58**, 199, 222.

CHANDRASEKHAR, S., 1943, *Rev. Mod. Phys.*, **15**, 2; 1949, *Ibid.*, **21**, 383.

DALY, E. F., 1947, *Thesis*, University of Cambridge (quoted by Fellgett 1949 a).

DAVENPORT, W. B., Jr., JOHNSON, R. A., and MIDDLETON, D., 1952, *J. Appl. Phys.*, **23**, 377.

EINSTEIN, A., 1905, *Ann. Phys., Lpz.*, **17**, 549; 1906, *Ibid.*, **19**, 371.

FELLER, W., 1950, *Probability Theory and its Applications* (New York : Wiley).

FELLGETT, P. B., 1949 a, *Proc. Phys. Soc.* B, **62**, 351; 1949 b, *J. Opt. Soc. Amer.*, **39**, 970.

FIRESTONE, F. A., 1932, *Rev. Sci. Instrum.*, **3**, 163.

FRANKENHAEUSER, B., and MACDONALD, D. K. C., 1949, *J. Sci. Instrum.*, **26**, 145

FÜRTH, R., 1948, *Proc. Roy. Soc.* A, **192**, 593; 1950, *Phil. Mag.*, **41**, 1255.

GISOLF, J. H., 1949, *Physica*, **15**, 825.

GOLAY, M. J. E., 1947, *Rev. Sci. Instrum.*, **18**, 347; 1949, *Ibid.*, **20**, 816.

GOLDMAN, S., 1948, *Frequency Analysis, Modulation and Noise* (New York : McGraw-Hill).

GREEN, M. S., 1951, *J. Chem. Phys.*, **19**, 1036.

GREENE, R. F., and CALLEN, H. B., 1951, *Phys. Rev.*, **83**, 1231.

HARDY, J. D., 1930, *Rev. Sci. Instrum.*, **1**, 429.

DEN HARTOG, H., and MULLER, F. A., 1947, *Physica*, **13**, 571.

HILL, A. V., 1934, *J. Sci. Instrum.*, **11**, 309; 1938, *Proc. Roy. Soc.* B, **126**, 136; 1948, *J. Sci. Instrum.*, **25**, 225; 1949, *Nature, Lond.*, **163**, 994.

ISING, G., 1926, *Phil. Mag.*, **1**, 827; 1932, *Ann. Phys., Lpz.*, **14**, 755.

JACKSON, J. L., 1952, *Phys. Rev.*, **87**, 471.

JOHNSON, J. B., 1928, *Phys. Rev.*, **32**, 97.

JONES, R. CLARK, 1947, *J. Opt. Soc. Amer.*, **37**, 879.

JONES, R. V., 1951, *Proc. Phys. Soc.* B, **64**, 469.

JONES, R. V., and MCCOMBIE, C. W., 1952, *Phil. Trans. Roy. Soc.* A, **244**, 205.

KAPPLER, E., 1931, *Ann. Phys., Lpz.*, **11**, 233; 1938, *Ibid.*, **31**, 619.

KHINTCHINE, A., 1934, *Math. Ann.*, **109**, 604.

KIRKWOOD, J. G., 1946, *J. Chem. Phys.*, **14**, 180.

KRAMERS, H. A., 1940, *Physica*, **7**, 284.

LAWSON, J. L., and UHLENBECK, G. E., 1950, *Threshold Signals* (New York : McGraw-Hill).

LEWIS, W. B., 1947, *Proc. Phys. Soc.*, **59**, 34.

MACDONALD, D. K. C., 1948, *Rep. Prog. Phys.*, **12**, 56 (London : Physical Society); 1950, *Phil. Mag.*, **41**, 814.

MACFARLANE, G. G., 1947, *Proc. Phys. Soc.*, **59**, 366; 1950, *Ibid.* B, **63**, 807.

MIDDLETON, D., 1951, *J. Appl. Phys.*, **22**, 1143, 1153.

MILATZ, J. M. W., and VAN DER VELDEN, H. A., 1943, *Physica*, **10**, 369.

MOLL, W. J. H., and BURGER, H. C., 1925, *Phil. Mag.*, **50**, 626.

MOSS, E. B., 1935, *J. Sci. Instrum.*, **12**, 141.

MOSS, T. S., 1950, *J. Opt. Soc. Amer.*, **40**, 603.

MOULLIN, E. B., 1938, *Spontaneous Fluctuations of Voltage* (Oxford : Clarendon Press), p. 216.

MOYAL, J. E., 1949, *J. R. Statist. Soc.* B, **11**, 151.

MÜSER, H., 1951, *Z. Phys.*, **129**, 504.

NYQUIST, H., 1928, *Phys. Rev.*, **32**, 110.

ORNSTEIN, L. S., BURGER, H. C., TAYLOR, J., and CLARKSON, W., 1927, *Proc. Roy. Soc.* A, **115**, 391.

PASSOTH, G., 1942, *Ann. Phys., Lpz.*, **42**, 49.

PERRIN, J., 1909, *Ann. Chim. Phys.* (Ser. 8), **18**, 5.

DU PRÉ, F. K., 1950, *Phys. Rev.*, **78**, 615.

PRESTON, J. S., 1946, *J. Sci. Instrum.*, **23**, 173.

PUMPER, E. YA., 1949. *Dokl. Akad. Nauk, SSSR*, **66**, 41.

RICE, S. O., 1944, *Bell. Syst. Tech. J.*, **23**, 282; 1945, *Ibid.*, **24**, 46.
RICHARDSON, J. M., 1950, *Bell. Syst. Tech. J.*, **29**, 117.
SCHMITT, O. H., 1938, *J. Sci. Instrum.*, **15**, 234.
SCHOTTKY, W., 1926, *Phys. Rev.*, **28**, 74.
SCHREMP, E. J., 1948, *Vacuum Tube Amplifiers* (New York : McGraw-Hill), p. 496.
SCHROEDINGER, E., 1946, *Statistical Thermodynamics* (Cambridge : University Press), p. 25.
SHANNON, C. E., 1948, *Bell. Syst. Tech. J.*, **27**, 379, 623.
SMITH, R. A., 1951, *Telecommunications Research Establishment Journal*, July, 1951, 1.
SPROULL, R. L., 1945, *Phys. Rev.*, **67**, 166.
STRUTT, M. J. O., and VAN DER ZIEL, A., 1942, *Physica*, **9**, 513, 528.
SURDIN, M., 1949, *J. Phys. Radium*, **10**, 253; 1951, *Physica*, **17**, 548.
TOLMAN, R. C., 1938, *Principles of Statistical Mechanics* (Oxford : University Press), p. 93.
TWISS, R. Q., and BEERS, Y., 1948, *Vacuum Tube Amplifiers* (New York : McGraw-Hill), p. 615.
UHLENBECK, G. E., and GOUDSMIT, S., 1929, *Phys. Rev.*, **34**, 145.
USPENSKY, T. V., 1937, *Introduction to Mathematical Probability Theory* (New York: McGraw-Hill).
VAN LEAR, G. A., Jr., 1933, *Rev. Sci. Instrum.*, **4**, 21.
VAN ZOLINGEN, J. J., 1948, *Dissertation*, Leiden.
WANG, M. C., and UHLENBECK, G. E., 1945, *Rev. Mod. Phys.*, **17**, 323.
WIENER, N., 1930, *Act. Math., Stockholm*, **55**, 117; 1949, *Extrapolation, Interpolation and Smoothing of Stationary Time Series* (New York : Wiley; London : Chapman and Hall), p. 81.
WOODWARD, P. M., 1948, *Telecommunications Research Establishment Journal*, Oct. 1948, 1.
WOODWARD, P. M., and DAVIES, I. L., 1952, *Proc. Instn. Elect. Engrs.*, **99**, 37.
ZERNIKE, F., 1926, *Z. Phys.*, **40**, 628; 1932, *Ibid.*, **79**, 516.
VAN DER ZIEL, A., 1942, *Physica*, **9**, 177; 1950, *Ibid.*, **16**, 359; 1952, *Advances in Electronics* (New York: Academic Press), p. 109.

COSMOLOGY

By W. H. McCREA
Royal Holloway College, University of London, and
Gonville and Caius College, Cambridge

CONTENTS

		PAGE
§ 1.	Introduction	322
§ 2.	Nature and aims of cosmology	323
§ 3.	The observed universe	324
§ 4.	Model universes	326
4.1.	Smoothed-out universe	326
4.2.	Cosmological principle	326
4.3.	Simple model	327
4.4.	Space–time	330
4.5.	Methodology	331
§ 5.	Kinematic relativity	332
5.1.	Milne's hydrodynamic system	332
5.2.	Simplified presentation	333
5.3.	Discussion	334
5.4.	Further development of kinematic relativity	335
5.5.	Milne's theory of gravitation	337
§ 6.	Steady-state theory	339
6.1.	General approach	339
6.2.	Bondi and Gold's approach	340
6.3.	Perfect cosmological principle	341
6.4.	Simplified presentation	341
6.5.	Creation	342
6.6.	Space–time	343
§ 7.	Newtonian universes	343
7.1.	Extrapolatory theory	343
7.2.	Newtonian universe	343
7.3.	Remarks	345
7.4.	Generalized newtonian models	346
7.5.	Particular models	347
§ 8.	Relativistic universes	350
8.1.	Comparison with newtonian theory	350
8.2.	Observable relations	351
§ 9.	Hoyle's theory	353
§ 10.	Physical discussion	353
10.1.	Preliminary results	354
10.2.	Derived results	355
10.3.	Conclusions	358
§ 11.	Creation of matter	358
11.1.	Unique creation	358
11.2.	Continual creation	359
§ 12.	Galaxies	359
§ 13.	Appendix on current developments	361
	References	362

ABSTRACT. The choice of treatment in this review is influenced by the nature of a number of other existing reviews of cosmology. Observational results are here summarized only briefly because, for the most part, the results obtained after the appearance of these other reviews have not yet been published. Cosmological theories fall into two main categories according as they start from postulates concerning the universe as a whole and seek to deduce physical laws, or start from postulated laws and construct model universes compatible with such laws. Of the cosmologies here reviewed, those of kinematic relativity and of the steady-state theory belong to the first category, while those of newtonian theory and of general relativity theory and its extensions belong to the second. To a considerable extent it proves possible to imitate the procedures of all these theories for the case of material moving in classical space–time. In this way a largely unified presentation is obtained and comparison of the theories is facilitated. At the same time the standard procedures of the theories are also indicated in the review. In comparing the theoretical models with the actual universe emphasis is laid upon the broad aspects about which there is a considerable measure of agreement. The possibilities that have been investigated of relating the theories to the problem of the origin of the chemical elements and to the problem of the origin and evolution of the galaxies are mentioned. The whole subject is in a somewhat inconclusive state, and the main purpose of the review is to describe the lines of thought being pursued in current work.

§1. INTRODUCTION

THE bulk of existing work on modern cosmology is so immense that it would be ridiculous to try to 'report' upon it in the way in which a writer would be expected to report upon some new and highly specialized branch of physics. In any case, cosmology is not established as a part of physics in the usual sense. But it does at least provide some impressive applications of ordinary physics. Also it may be expected to give some insight into the origin and status of the laws of physics. Presumably it is for such reasons that a 'report' upon cosmology is called for in the present series. The form of the report should be determined accordingly.

Other considerations arise from there being in the field at the present time, as is well known, several apparently rival theories about which there has been considerable controversy. It is not a trivial matter to exhibit the essentials in which these theories differ and those in which they agree. Indeed, the controversy has been occasioned more by this difficulty than by the differences themselves. One aim of this paper is to try to clarify the situation by giving, so far as is possible in a simple fashion, a unified presentation of the main theories.

The modern study of cosmology may be dated from 1917, when Einstein and de Sitter published their relativistic models of the universe. Enormous effort has gone into the development of the subject since then, but particularly since about 1930, when Hubble, following upon earlier observational work by Slipher, had accumulated enough observations to announce his law of the apparent recession of the nebulae, and when Eddington called attention to the significance of earlier theoretical work by Friedmann and Lemaître on non-static models. From time to time leading contributors have reviewed the progress of the work. Einstein (1951) and de Sitter (1932) have done so. Robertson (1933) reviewed it from the standpoint of general relativity theory using a very fundamental mathematical approach. Hubble (1936, 1937) reviewed it from the observational standpoint. Heckmann (1942) and Bondi (1948, 1952) have critically reviewed all the main theories. Theories of the origin of the chemical elements have been reviewed by ter Haar (1950) and Alpher and Hermann (1950). There are also well-known monographs on particular theories by Eddington (1936, 1946),

Jordan (1947, 1952) Lemaître (1950), and Milne (1935, 1948, 1952). A list of some other books is given at the end of the references.

The existence of Bondi's (1952) very recent and comprehensive review means that there would be little value in giving here a more condensed treatment along the same lines. The form of the present paper has to a large extent been determined by this consideration. Frequent reference is made to Bondi's book, which can be consulted also on all matters that should be found in a systematic survey but are not included in the present treatment. Among such matters are a general survey of the observational evidence, a summary of the actual mathematics of the theories (the present discussion being concerned with their aims rather than their techniques), and some sketch of the theories of Eddington, Dirac and Jordan (these having so far not greatly influenced the general trend of cosmology). Finally, no account of the history of the subject is attempted here.

§2. NATURE AND AIMS OF COSMOLOGY

Cosmology may be said to be that part of physics for which the main interest lies in what exists 'actually' or 'accidentally'. It may be contrasted with ordinary physics, for which the main interest lies in what exists 'typically' or 'generally'.

The ordinary physicist is interested in, for example, the properties of iron. He is interested because iron *exists*. But his interest, as an ordinary physicist, does not depend upon where and when and in what quantities iron exists. The latter features are merely 'accidental' from his standpoint. On the other hand, whatever interest the geophysicist has in iron depends essentially upon where, and in what quantities, iron exists in the Earth. That is, he is interested in the iron that 'actually' exists. The geophysicist could therefore be called a cosmologist and, incidentally, it is nowadays a known practice to include geophysics under the head of 'cosmical physics'.

To take another example, the solar physicist is interested in the Sun that actually exists. The endless observations he makes of its behaviour have no immediate counterpart in the doings of a laboratory physicist, who would never show the same interest in any particular system.

But consider now how the solar physicist proceeds if he wants to work out, say, the internal constitution of the Sun. He formulates a system of differential equations expressing what he believes to be the correct relation between the significant parameters. This system has infinitely many solutions. By using suitable boundary conditions, such as that the total mass should be equal to the observed mass, he picks out the solution which he proposes to compare with the actual Sun. The points to notice are that he would regard all the other solutions (at any rate within a certain range) as representing suns whose existence is equally possible, and that he regards the boundary conditions used as 'accidental' to the particular solution selected. Thus he converts his problem into one of ordinary physics.

Pursuing such ideas, it begins to look as though in practice there is no physics different from ordinary physics, and that physics never comes to grip with 'actual' existence.

There is, however, one system for which the 'actual' existence is all-important and whose study can scarcely be converted into a problem of ordinary physics. That system is whatever we mean by the *universe*. For it seems to be meaningless to describe the universe as belonging to any class of entities other than itself that 'could exist'. When we speak of the other solutions of the equations of

stellar structure, besides the one we are interested in at the moment, as representing systems that could exist, we mean that they could exist in the universe as we know it. Clearly no such attitude of thought is possible towards the universe itself.

Cosmology is usually described as the study of the universe as a whole or of the large-scale properties of the universe. We seem now to have discovered the reason why this description is suitable. It appears that at some stage in extending the scale of physics there comes a real difference from ordinary physics; at that stage we reach cosmology.

It might be suggested that cosmology should therefore not be regarded as part of physics, but this is merely a matter of definition. In any case, to restrict physics in such a way would be to destroy the meaning of the well-established concept of 'the physical universe'.

These considerations are mentioned in order to show that cosmology is not merely an application of physics in a particular field, and that cosmology may be expected to include some sets of ideas other than those of ordinary physics.

Nevertheless, it is not profitable to pursue these considerations very far. It is of more value to inspect the attempts that have actually been made to study the universe in the large.

The immediate aims of these attempts are to discover the large-scale properties of the universe, such as the distribution, motion and chemical composition of matter and to establish relations between them. But it should be added that the study of cosmology has sometimes arisen out of the study of problems, such as that of the relation between gravitation and inertia or those of the fundamental constants of physics, which do not at first sight necessarily involve the large-scale properties of the universe. These are, however, problems of what 'actually' exists, and so, from our point of view, they are cosmological from the outset (e.g. Sciama 1953).

§ 3. THE OBSERVED UNIVERSE

On the scale on which the universe is studied in cosmology as such it is customary to regard the *galaxies* as the natural units into which the contents of the universe are organized. The galaxies are systems similar, in a very broad sense, to our own Milky Way Galaxy. The term is equivalent to 'extra-galactic nebulae' and includes the 'spiral nebulae'.

The customary view of the status of the galaxies rests upon the fact that all the *observed* material in the universe can be assigned as 'belonging' definitely and unambiguously to one particular galaxy or another. (Any exceptions to this statement are at present still dubious and, for our purposes, altogether negligible; the possible presence of hitherto unobserved material will be mentioned later).

It is well known that the galaxies differ greatly in absolute luminosity, in dimensions, in the amount of detectable interstellar 'dust' they contain, in the nature of their stellar populations, and particularly in their shape and structure. Nevertheless, the fact that they can be arranged according to these characteristics in a well-ordered (though actually branching) linear sequence confirms the view that they are systems in a single category.

The galaxies are not all distributed over the sky as a random distribution of individuals. Many, known as 'field galaxies' (formerly estimated at about

80 per cent of all observed galaxies, but see Shane and Wirtanen 1950), are apparently so distributed. The rest are distributed in 'clusters' (which are now sometimes sub-classified further) comprising something of the order of ten galaxies at one extreme and something of the order of a thousand at the other. The facts that there are so many field galaxies and that there does not seem to be evidence for the existence of 'clusters of clusters' support the view generally held that the observed galaxies do not belong to still greater units in the organization of the universe. (It might be held that there is nothing to confirm or deny the the view that all these galaxies are part of just one such unit, the universe being possibly an indefinitely continued hierarchy of systems. But the law of red-shifts (see below), if it holds for the whole observable universe, sets a limit to the extent of this universe and seems to render such a concept meaningless.)

For a summary of most of the existing published observational results reference should still be made to Hubble's monographs (1936, 1937). To a great extent the enlargement of the already vast amount of material there surveyed has had to await the bringing into operation of the new equipment on Mt. Palomar. Hubble (1951) has recently indicated the scope of the new work.

Only a few general results, in round figures, need be quoted here. Most of the galaxies are estimated to lie in the range of total luminosity of 10^7 to 10^9 suns, and in the range of 10^9 to 10^{11} solar masses. Within a sphere of radius of about 5×10^8 light-years (the estimated average distance of the faintest galaxies detectable with the 100-inch telescope at Mt. Wilson) there are estimated to be about 10^8 galaxies. This is usually taken to correspond to a mean density of matter throughout this region of about 10^{-30} g cm^{-3}. The distribution of the galaxies in the region does show considerable irregularities, but Hubble's conclusion is that it can be interpreted as being ***homogeneous and isotropic*** in the large.

The most famous characteristic of the whole assembly of galaxies is its apparent state of ***expansion***. The spectra of the galaxies show red-shifts of the identifiable spectral lines which increase with distance. If interpreted as doppler-shifts they lead to the result

$$\text{speed of recession} = \text{distance} \times T^{-1}. \qquad \ldots\ldots(3.1)$$

This is known as *Hubble's law* and T^{-1} as Hubble's constant. T is of the dimensions of time, and the value derived from the observations is given as

$$T = 1{\cdot}8 \times 10^9 \text{ years}. \qquad \ldots\ldots(3.2)$$

The measurements concerned have been extended to galaxies out to about $2{\cdot}5 \times 10^8$ light years. In rough figures, Hubble's law states that the apparent speed of recession is 100 miles per second per million light years distance. The apparent homogeneity of the universe leads to the inference that an observer anywhere in it would witness the same apparent recession (see § 13).

In addition to this systematic motion, the galaxies are judged to have random peculiar motions of the order of 100–200 miles per second.

As regards more recent observations, perhaps it is sufficient here to mention only two results. One is that the intrinsic properties of galaxies vary over ranges greater than had previously been estimated (Baade 1951). The other is that observations by Stebbins and Whitford (1948) seem to mean that, at any rate for certain classes of galaxy, the *continuous* spectrum shows a reddening with distance, i.e. that it contains a greater proportion of red light for a distant galaxy than for

a relatively near one, and that the effect significantly exceeds the amount that would be deduced from Hubble's law. There is no generally accepted explanation of this effect. But, if it comes to be established for galaxies in general, it will necessitate some revision of the quantitative interpretation of nearly all the other observations.

§4. MODEL UNIVERSES

4.1. *Smoothed-out Universe*

The aim of cosmology is to produce a theoretical model of the universe whose theoretical behaviour reproduces the actual behaviour of the actual universe. More accurately, we say that the 'observable relations' for the model should be in agreement with experience; we return to this point later. It will be convenient to use 'universe' to mean whatever model is being discussed. (Some writers appear to use the term 'model' only in connection with a theory belonging to a particular category; no such restriction will be implied here.)

We are about to discuss theories of the universe on the largest scale on which we can study it. All such theories seek only to form a model of what the actual universe would be like if its contents are supposed to be *smoothed out* so as to behave like a continuous fluid.

4.2. *Cosmological Principle*

All the theories to be discussed require their models to conform to the *cosmological principle* (CP), though we shall see that they do so for somewhat different reasons.

We suppose there to be an observer permanently attached to each particle of the fluid: unless otherwise stated, 'observer' means such a 'fundamental observer'. Then the CP states:

(A) *Every observer is equivalent to every other.*

That is to say, every observer is assumed to see the same sequence of 'world-pictures'.

If the universe is not in a steady state the observers can be supposed each to assign the same epoch t to the same world-picture; we call t the *cosmic time*. In other words, all observers may use the universe itself as a 'cosmic clock'.

Any monotonic increasing function of t serves equally well as cosmic time. But in general a particular time-scale is determined by some 'local clock' in the possession of each observer. By virtue of the CP all such clocks must be regarded as equivalent. So we can say that the observers may adjust the zeros of their local clocks so that they read cosmic time on this particular scale. Any two observers then have the same world-picture when their clocks read the same time. We reserve the symbol t for time kept in this way.

We have been using the term 'world-picture' for what an observer sees at any particular instant. He does not, of course, see any other region of the universe as it is at the cosmic epoch at which he makes his observation, but as it was at some earlier epoch depending on the finite time of light-travel from that region to himself. But we are ordinarily accustomed to think that we may speak of what a system would be like if we could see it 'all at the same time'. Actually the theory of relativity in general denies a meaning to this description since it denies a meaning to absolute simultaneity. However, the concept of cosmic time restores a meaning to absolute simultaneity for the particular systems we are studying.

So we can give some meaning to what a model universe would be like if an observer could see it all as it is at the same cosmic epoch t. This we shall call merely the model 'at epoch t'. Any specification of it we shall call a 'world-map'. The expressions 'world-picture' and 'world-map' are due to Milne. (The ideas we are here trying briefly to describe are legitimate for their immediate purpose. But they have been criticized as not being capable of extension to a system more like the actual universe which does not conform strictly to the CP (Bondi 1952, p. 71).)

Clearly the CP requires the model at epoch t to be *homogeneous*. In the form stated it does not require the model to be isotropic. However, all theories do in fact make the additional assumption:

The universe is isotropic at every cosmic epoch.

Henceforth we take this to be included in the CP.

Now Walker (1944 a, b, c) has pointed out that all the requirements of the CP are satisfied if we assume:

(A') *The universe is always isotropic as seen by any fundamental observer.*

The converse is true for the CP as extended by the assumption of isotropy. But, unlike Walker's, this latter assumption is not stated in terms of what an observer actually sees. Walker's demonstration applies to a very general treatment of space–time which includes all the treatments used in the theories considered here.

Remarks. The form (A') of the CP provides a very simple partial test as to whether the actual universe satisfies the principle. For, using comparatively simple observations, we ought to be able to decide whether the universe appears to be isotropic about ourselves. If it does, and if we then assume only that there is nothing special about ourselves, it would follow from Walker's considerations that the actual universe satisfies the CP. Also it would follow that 'we', i.e. an observer on our own Galaxy, rank as a fundamental observer. The test would be incomplete because it would depend upon the assumption of the non-speciality of ourselves.

It is probably difficult to believe that the universe should be other than homogeneous and isotropic in the large. (Of course it is easy to construct theoretical models which do not possess these properties, but most of them appear obviously unnatural.) There is, however, one way in particular in which the universe could fail to possess these properties as interpreted here. That is if the imagined smoothing out of the contents of the universe were found to be meaningless. This could happen if the average density throughout a region tends to zero as the volume of the region tends to infinity. In that case much of the present discussion would fail to apply. This would be a possibility, for example, were the universe of the 'hierarchical' type (Bondi 1952, pp. 14, 19). The contents of such a universe could be such that they are not representable in any intelligible way as a continuous fluid.

However, as we have said, we are here reviewing theoretical models which do satisfy the CP.

4.3. *Simple Model*

It will prove of great assistance in the subsequent exposition to consider first the following 'simple' model. The model itself is not intrinsically simpler than the others which will be considered, but it enables us to give a simplified ***presentation*** of the whole subject.

Let the space of the model be euclidean and let the time t be universal newtonian time. Let there be material of density ρ, not everywhere zero, moving as a continuous fluid. By this we mean that the velocity of the fluid is defined in the standard way of hydrodynamics and that its time variation is that produced by the actual motion of particles of the fluid. As before, suppose there is a (fundamental) observer permanently attached to each particle of the fluid.

The specified nature of space and time implies the 'classical' composition of velocities. For the present, 'density' is taken to refer to any particular measure of quantity of material. We suppose it to have a meaning to assert that all observers use a common system of units. The various functions to be employed are assumed to satisfy any required conditions of integrability or differentiability.

We now assume that the model satisfies the CP in the form (A').

The following deductions can be made:

(*a*) The density ρ is uniform at each epoch and so can be a function $\rho(t)$ of the epoch only. It follows that at any epoch there is one, and only one, observer at every point of the space. Any other property of the system specified by a scalar quantity is similarly a function of t only.

(*b*) Any observer O can choose a frame of reference F with origin at himself so that, relative to this frame, the motion of the fluid is purely radial with velocity depending at most upon radial distance and epoch. Thus, if $\mathbf{V}$ is the velocity at the point having position vector $\mathbf{q}$ in F, then

$$\mathbf{V} = V(q, t)\mathbf{q} \qquad \ldots\ldots(4.1)$$

where V is a scalar function of at most $q(=|\mathbf{q}|)$ and t.

(*c*) The function $V(q, t)$ is independent of q and is the same function $V(t)$ for all observers. So in place of (4.1) we have

$$\mathbf{V} = V(t)\mathbf{q} \qquad \ldots\ldots(4.2)$$

valid for every observer using his own frame corresponding to F, where $\mathbf{V}$, $\mathbf{q}$ are relative to the frame used.

(*d*) The frames such as F for any two observers do not rotate relative to each other.

To sum up: *Every observer regards the fluid as having uniform density and as receding from or approaching himself with speed proportional to the distance from him, the factor of proportionality and the density being functions of the epoch, the same for all observers.*

The derivation need not be given in detail; that of (*a*) is almost immediate; (*b*) follows from the assumed isotropy about O; (*c*), (*d*) then follow from the assumed isotropy about each of an arbitrary pair of observers.

It is clear that, conversely, for any pair of functions $\rho(t)$, $V(t)$ the model satisfies the definition and the CP. Thus the model we have been considering does *exist* in the mathematical sense.

From now on any equations we use for the model will be valid for any observer O using his F-frame.

Quantities $R(t)$, T, $Q(t)$, $M(t)$. We now define the function $R(t)$ as the solution of the differential equation

$$R'/R = V \qquad \ldots\ldots(4.3)$$

for which

$$R_0 \equiv R(t_0) = 1 \qquad \ldots\ldots(4.4)$$

where t_0 is some particular value of t. Thus

$$R(t) = \exp \int_{t_0}^{t} V dt. \qquad \ldots\ldots(4.5)$$

For any particle of the fluid we have from (4.2)

$$dq/dt = qV(t). \qquad \ldots\ldots(4.6)$$

Hence, using (4.3), we have

$$q = rR(t) \qquad \ldots\ldots(4.7)$$

$$dq/dt = rR'(t). \qquad \ldots\ldots(4.8)$$

Here $r(\geqslant 0)$ is a constant of integration that permanently characterizes the particle considered, being in fact its distance $q(t_0)$ from O at time t_0.

It is worth noting that the form of q as a function of t in (4.7) automatically ensures that the particles of the fluid never overrun each other during the motion.

We define the constant T having the dimensions of time by

$$V_0 \equiv V(t_0) = 1/T. \qquad \ldots\ldots(4.9)$$

If c is some particular speed which will be taken to be the speed of light, then for any value of t for which $V(t) > 0$ there exists a value $Q(t)$ of q such that

$$Q(t) \,.\, V(t) = c. \qquad \ldots\ldots(4.10)$$

The amount of material within distance $Q(t)$ is $M(t)$, where

$$M(t) = \tfrac{4}{3}\pi c^3 \rho(t)[V(t)]^{-3}. \qquad \ldots\ldots(4.11)$$

S-model. It will be convenient to refer to the present model as the S-model. A particular S-model is completely specified by the functions $\rho(t)$, $V(t)$ or by $\rho(t)$, $R(t)$.

Time-keeping. The following remarks will later help to clarify a feature of Milne's cosmology. If $R(t)$ is not constant for all t, there must exist intervals in which it is increasing or decreasing. Consider one in which it is increasing and write

$$t^* = t_0^* R(t) \quad (t_0^* = \text{constant}). \qquad \ldots\ldots(4.12)$$

In the interval t^* is then in 1–1 correspondence with t (and, in general, the origin of t^* is outside the interval).

From (4.7) we should then have

$$q = (r/t_0^*)t^* \qquad \ldots\ldots(4.13)$$

This gives

$$dq/dt^* = r/t_0^* \qquad \ldots\ldots(4.14)$$

and so we can write $dq/dt^* = qV(t^*)$, where $V(t^*) = 1/t^*$. $\ldots\ldots(4.15)$

In the interval every observer could use t^* instead of t as a measure of time. In t^*-time (4.14) shows that every particle of the fluid recedes from the observer with *uniform* velocity, and (4.15) shows that at epoch t^* the particles at given distance q have speed of recession inversely proportional to t^*.

As we have already said, the observers could use the model itself as a 'cosmic' clock and in particular they could use it to keep t^*-time. If the observers do this, and have no other means of time-keeping, then every expanding model is equivalent to every other so far as the motion is concerned. An expanding S-model would then be completely specified by a single function $\rho(t^*)$, giving the density at epoch t^*.

Corresponding remarks apply to contracting systems.

However, in taking t to be newtonian time we have to suppose the observers to be equipped with some form of local clocks which keep this time. We shall retain this supposition in regard to the S-models, so that such models are not, in fact, observationally equivalent if the functions $R(t)$ are different.

4.4. *Space–Time*

Light. In constructing a model we ought to have regard to the propagation of light since this conditions the observations made by the observers.

The first thing to notice is that the postulated equivalence of the observers implies that the speed of light must be the same for all of them. But this at once exposes the deficiency of our S-model. For it is constructed with the aid of classical kinematics, which does not admit the possibility of light having the same speed for all observers unless they are at rest or unless the speed of light is infinite.

It is well known that the theory of general relativity looks after this feature by treating space–time as a Riemannian four-space and postulating that the world-lines of light-signals are null-geodesics in this space.

We shall therefore consider the space–time whose metric is

$$ds^2 = c^2\,dt^2 - R^2(dr^2 + r^2\,d\theta^2 + r^2\sin^2\theta\,d\phi^2), \qquad \ldots\ldots(4.16)$$

where $R \equiv R(t)$ is a function of t only. This is like the well-known Minkowski space–time of special relativity except that it has Rr in place of r and $R\,dr$ in place of dr. So for any given value of t we could regard Rr as the distance from the origin of the point with coordinates r, θ, ϕ. If, therefore, we suppose there to be an observer corresponding to every fixed set of values of r, θ, ϕ and treat t as the time, then the relative motion of the observers would be like that in the S-model, with r, R playing the same parts as before. There is one of these observers for which $r = 0$. But, by analogy with a transformation of polar coordinates in ordinary space, if we give any other observer this role the form of (4.16) is unaltered. So all the observers are verified to be equivalent. We could then add the assumption that the observers are moving with a fluid of density $\rho(t)$.

Thus we should have a system giving a world-map formally the same as before. But by associating it with the space–time (4.16) and applying the relativistic treatment of light-propagation we should remedy the deficiency in the S-model.

We then notice that we may have restricted the possibilities of our S-model by taking its space to be euclidean. But it is evident that we could not have taken it to be more general than a space of constant curvature. Otherwise not every part would be equivalent to every other and the CP could not be satisfied.

The way to incorporate this possibility into (4.16) is to make the sub-space $t = \text{const.}$ a space of constant curvature, though this curvature could depend upon t. From elementary differential geometry it is known that this would be achieved if we replace (4.16) by

$$ds^2 = c^2\,dt^2 - \frac{R^2}{(1 + \frac{1}{4}kr^2)^2}(dr^2 + r^2\,d\theta^2 + r^2\sin^2\theta\,d\phi^2). \qquad \ldots\ldots(4.17)$$

Then the sub-space $t = \text{const.}$ has constant curvature $\pm\,|k|^{1/2}/R(t)$, according as k is positive or negative.

It can be verified as before that this does in fact lead in the same way to a model satisfying the CP.

G-model. The foregoing is not a proper derivation of (4.17). The general result we have to quote is that a set of requirements equivalent to those leading to our S-model (but without any assumption about a particular kind of geometry) leads to a model that can be described by a Riemannian space–time of the form (4.17) in which the propagation of light is to be treated in the same way as in general relativity and in which every world-line along which r, θ, ϕ are constant is the world-line of an observer. The observers move with a fluid whose density as determined by any observer for the fluid in his own neighbourhood is $\rho(t)$, a function of t only. The variable t plays the part of cosmic time in the model. We call this the G-model (general model).

The postulates leading to this result are so general that they are found to be common to all the theories discussed below, and each theory does, in fact, possess a stage in its development at which it produces this result. Some reservation is called for if we include the newtonian theory on its own account, but we shall see that this theory passes directly into the relativistic one when the full result is used.

The most general derivation of (4.17) is due to Robertson (1935, 1936) and to Walker (1936). It ought to be emphasized that no equations of motion, no field equations and no theory of radiation are required. All that is postulated about 'light' is that there is some means of observing or signalling that does not violate the CP. Admittedly, it is subsequently assumed that what we know physically as radiation has the characteristic property of this 'light' (see also Weyl 1923).

A G-model is again determined by the functions $\rho(t)$, $R(t)$ along now with a disposable constant k. The G-model with $k=0$ may be thought of as the S-model adjusted to an adequate treatment of light-propagation. A G-model with $k\neq 0$ may in turn be thought of as an adjusted S-model based upon a modified geometry, and it can be remembered that this modification is small over sufficiently small distances.

It turns out that all the essentials of all the theories can be explained in terms of the S-model. We propose to explain them in this way in the hope that it will assist the comparison of these essentials in the various cases. The passage to the G-model can be viewed as not a great deal more than a tidying up of the mathematical treatment.

4.5. *Methodology*

At this stage we have to ask, do the contents of the universe and their behaviour determine, or are these determined by, the laws of physics?

This is not meant to be taken as a critically formulated question. It is stated in order to indicate a dilemma in which the theoretical cosmologist finds himself placed.

The laboratory physicist is not troubled by the problem in any practical way. The amount of the universe that he can vary in making his experiments in order to discover the laws of physics is so small that he can safely think of the laws as something fixed. Moreover, he regards these laws essentially as a means of predicting the results of other experiments.

The cosmologist, however, professes to deal with the whole universe. It would certainly be absurd for him to think of the laws of physics as something fixed independently of the system he is dealing with. But in any case he cannot vary the universe in order to discover physical laws as an experimental physicist

does. Also there is no other universe whose behaviour the laws are required to predict. It appears therefore as though laws of physics in the usual sense are scarcely relevant to the study of the universe as a whole.

The suggestion is, then, that the cosmologist must start with postulates concerning the universe and seek to deduce from them some further properties of the universe. In the course of the work he may hope to deduce also something about the laws of physics, but he will not expect these laws to tell him anything about the universe as a whole. Perhaps the clearest statement of this attitude is Milne's (1952, p. 70). He wrote "the distribution of matter-in-motion in the universe controls and fixes the so-called laws of nature that the distribution obeys". Bondi (1952, p. 5) calls this the 'deductive' attitude to cosmology.

Following a different approach, the cosmologist can argue on the other hand that he would have no hesitation in applying 'known' laws of physics to predict properties of say, stars, the Galaxy, clusters of galaxies. Where then is this process of extrapolation to end, and need it end short of the whole universe?

In that case the suggestion is that the cosmologist should postulate the laws of physics and attempt to find a system that satisfies them, as he might say, completely, i.e. without any boundary conditions not determined by the system itself. Such a system could legitimately be called a universe. The cosmologist would hope to infer information about the general behaviour of the actual universe. This is what Bondi calls the 'extrapolating' attitude; it might also be called the 'astrophysical' attitude.

These are rough descriptions of the two main possible attitudes to the subject. One or other may be adopted on account of the worker's philosophy of nature or merely because he has to start somewhere. Much needless controversy has arisen through not appreciating the distinction between them and not recognizing which was being adopted in particular instances.

On account of the actual inseparability of the universe and its laws we cannot expect complete success from either procedure. We are prepared to find therefore that neither procedure is strictly adhered to in practice, and that any classification of theories is not very precise. With this reservation, we may call the cosmologies of kinematic relativity and of the steady-state theory 'deductive', and newtonian cosmology and relativistic cosmology 'extrapolatory'.

We proceed to a review of all the main current theories of cosmology and the model universes yielded by them. We take the deductive theories first largely because they call for more critical discussion.

§ 5. KINEMATIC RELATIVITY

5.1. *Milne's Hydrodynamic System*

Kinematic relativity was the creation of Milne (1935, 1948, 1952) and its cosmology was developed almost entirely by him. Sketched very briefly he claimed to proceed as follows: Milne gave 'operational' definitions of epoch and distance, first in one and then in three spatial dimensions. In these definitions the constancy of the speed of light is required only as a convention. He then defined what he called an 'equivalence' of observers and demonstrated its existence in the mathematical sense. On requiring this equivalence to conform to the CP it became a 'substratum' or 'hydrodynamic system', and this is the exact counterpart of our fluid accompanied by its fundamental observers.

Milne's time-keeping is here equivalent to using the universe as its own clock. Consequently, according to the argument given above (p. 329), his model is effectively unique as regards the motion of the fluid. He then claimed to get a unique expression for the density without making physical assumptions.

The resulting model Milne identified with the actual universe of galaxies, the particles of the fluid being taken to be the nuclei of the galaxies. The substratum is a continuous fluid in this model, as in any other we shall consider. But it was never altogether clear whether Milne thought of it as an immaterial matrix any idealized particle of which might or might not be 'occupied' by the nucleus of a galaxy, or whether he thought that the actual universe could be treated as a continuous distribution of galaxies. This point does not arise in other theories where the analogue of Milne's substratum is intended to represent all the contents of the universe when these are regarded as being smoothed-out over the space. Had Milne stopped at the derivation of his hydrodynamic model we should have taken it as having the same interpretation. But, since he proceeded to investigate his statistical models (see p. 336) consisting of the substratum *plus* something else, it is evident that his association of the galactic *nuclei* with his substratum meant that his own interpretation was something different.

In whatever sense the identification was intended, Milne did appear to claim that the actual universe *must* be like his model. For he considered that he had derived this model without making any physical hypotheses and had shown it to be unique. The argument was that there is only one universe that we can rationally conceive, and so the actual universe must be this one.

Milne's last book (1952) leaves little doubt that this was his view. He naturally regarded it as an extraordinary advance in physical thought. Though the criticisms about to be given seem to dispose of this claim, we shall see that great significance still attaches to some features of the theory.

5.2. *Simplified Presentation*

We construct an S-model to match Milne's hydrodynamic model as nearly as may be. If we seem to depart from the spirit of Milne's approach the reasons will become clear as we proceed.

As we know, an S-model is determined by the functions $R(t)$, $\rho(t)$, and we take these in turn.

Function $R(t)$. In Milne's model (*a*) the speed of light is the same for all observers at all epochs, (*b*) the observers move with uniform relative velocities.

Although our S-model itself does not deal satisfactorily with light-propagation, we know that, when this is remedied by going over to the G-model, the property (*a*) is certainly obtained. But if the observers measure the speed of light by any ordinary laboratory procedure and get a value independent of the epoch in one method of time-keeping they will not get such a result using any essentially different method of time-keeping. Hence (*a*) implies the use of some particular method of time-keeping. In constructing our S-model we must therefore assert:

A. The time-keeping to be used is implied from the outset. This is only a repetition of a condition previously assumed for S-models. Let t be the time concerned.

In order to match property (*b*) we have to postulate:

B. Every particle of the fluid moves with uniform non-zero velocity (relative to the F-frame of any other particle). This gives

$$V=1/t, \qquad R=t/t_0 \qquad \ldots\ldots(5.1)$$

and also

$$Q(t)=ct, \qquad T=t_0. \qquad \ldots\ldots(5.2)$$

Function $\rho(t)$. Milne took it for granted that the particles of his fluid have permanent existence. But our S-model could exist for any arbitrary function $\rho(t)$. Hence, in order to match Milne's requirements, we must postulate:

C. The material as described by $\rho(t)$ is conserved. The assumption is that the amount of material within any closed surface moving with the fluid is constant. From (4.7) the linear dimensions of such a surface are proportional to $R(t)$. Hence C gives

$$\rho(t)=\rho_0[R(t)]^{-3} \qquad \ldots\ldots(5.3)$$

where $\rho=\rho_0$ when $t=t_0$. Using (5.1) we get

$$\rho(t)=\rho_0 t_0^3/t^3. \qquad \ldots\ldots(5.4)$$

Also we note that (4.11), (5.2), (5.4) give

$$M(t)=\tfrac{4}{3}\pi c^3\rho_0 t_0^3. \qquad \ldots\ldots(5.5)$$

This S-model is very simple; it is merely a uniform fluid expanding at a constant rate in infinite euclidean space with conservation of material. The part of the system within distance $Q(t)$ of any observer gives a good representation of Milne's universe as mapped by the observer. However, we mention that Milne's universe is given exactly by the G-model with (Kermack and McCrea 1933)

$$R(t)=t/t_0, \qquad k=-1/c^2t_0^2. \qquad \ldots\ldots(5.6)$$

The results are collected in the table (p. 349).

5.3. *Discussion*

In order to get the particular S-model we had to make the physical hypotheses or postulates B, C and also the one (which we shall call D) implied by taking the space to be euclidean. Milne, on the other hand, claimed to have made no such hypotheses. It is important to explain this situation. What Milne did was virtually to re-discover the models corresponding to all $R(t)$. But he had an implicit assumption equivalent to requiring $R(t)$ to be monotonic increasing. Therefore he was able to transform all these models into the one with $R(t)\propto t$ by a change of time-keeping. So he regarded them all as physically equivalent. ***But he did not relate his time-keeping to any local laboratory observations that might be made by his observers***, in particular to their measurement of the speed of light. Something is required here which is not supplied by kinematic relativity at this or any later stage. When Milne transformed all the models into that having $R(t)\propto t$ he did so in such a way that he employed for the latter a definition of distance that preserved the constancy of the speed of light as measured in his new time and distance. But there is no purely logical reason for this distance to be the same as that measured locally by, say, a rigid scale, nor for it to be related to the rigid-scale distance in the same way for each of the different models Milne started from. The uniqueness claimed depends upon the implicit hypothesis that this is the case. Since Milne evidently wanted his distance to be ultimately identifiable locally with rigid-scale distance, the implied assumption is equivalent to B.

As regards C, there is no question that kinematic relativity contains this postulate. But it was scarcely recognized as a physical hypothesis in this connection until Bondi and Gold (1948) actually used a different hypothesis in its place.

As regards D, it has for some time been recognized (Bondi 1952, p. 129) that in his passage from one- to three-dimensional space Milne had made a hidden hypothesis which gave him a unique result. Actually this was equivalent to his arbitrarily selecting hyperbolic space rather than our euclidean space. But the point is that D does match an hypothesis in Milne's work.

Just as in our simplified presentation, Milne had therefore to make definite physical hypotheses in order to get his unique result. Without these his work would have given him the equivalent of the whole set of G-models for arbitrary $R(t)$, $\rho(t)$ and k.

5.4. *Further Development of Kinematic Relativity*

We have now to sketch Milne's further development of kinematic relativity so far as it concerns the cosmological problem. But for our purpose the exact forms of Milne's results are not essential. We can show their general character by imitating Milne's work, using the S-model obtained in (5·2) as a counterpart of Milne's hydrodynamic model. The results correspond to Milne's but are *not* the same and have no interest except as illustrating his procedure.

Free particle. The first step is to establish the equation of motion of a *free test-particle*. This is a particle supposed in general not to belong to the substratum but to be projected from an arbitrary point with an arbitrary velocity in the presence of the substratum and possibly of other free particles.

This is a new feature in the situation, and without new postulates nothing whatever can be said about it. Without stating them as such in his own presentation, Milne's additional postulates are:

E. There is an equation of motion for any free particle P, and this expresses its acceleration in terms of its velocity, position and the epoch.

Thus the equation has the form

$$d\mathbf{v}/dt = \mathbf{f}(\mathbf{v}, \mathbf{q}, t) \qquad \ldots\ldots(5.7)$$

where $\mathbf{v}$, $\mathbf{q}$ are the velocity and position of P, at epoch t, relative to the F-frame of any (fundamental) observer O.

F. The function $\mathbf{f}$ is the same whatever (fundamental) observer O is concerned.

G. The function $\mathbf{f}$ involves no dimensional constants.

Milne apparently considered that F is implied by the CP and that G had to be true because his system contained no dimensional constants. With regard to this, it must be remembered that he treated c as an arbitrarily chosen number. Even if c be regarded as a dimensional constant from our point of view we do not include it in $\mathbf{f}$ because we are here following Milne's procedure.

Let $\mathbf{v}, \mathbf{q}, t$ and $\mathbf{v}_1, \mathbf{q}_1, t_1$ apply to the same event in P's career relative to the frames F, F_1 of observers O, O_1 respectively. Let $\mathbf{a}, \mathbf{u}$ be the position and velocity of O_1 relative to F. In our S-system we then have $t_1 = t$, $\mathbf{q}_1 = \mathbf{q} - \mathbf{a}$, $\mathbf{v}_1 = \mathbf{v} - \mathbf{u}$, $\mathbf{u} = \mathbf{a}V$, the last equation being the appropriate case of (4.2). From these we can eliminate $\mathbf{a}, \mathbf{u}$ in only one way, obtaining

$$t_1 = t, \qquad \mathbf{q}_1 V - \mathbf{v}_1 = \mathbf{q}V - \mathbf{v}. \qquad \ldots\ldots(5.8)$$

Thus, in the parameters concerned, t and $\mathbf{q}V-\mathbf{v}$ are the only invariants. From E, F it follows that $\mathbf{f}$ must be a function of these. Moreover, it is by definition a vector function of dimensions length÷(time)^2. No dimensional constants being allowed by G, we have, therefore,

$$\mathbf{f}(\mathbf{v},\mathbf{q},t)\equiv K(\mathbf{q}V-\mathbf{v})/t \qquad \ldots\ldots(5.9)$$

when K is a purely numerical constant.

Using the value $V=1/t$ for the model (5.1), the equation of motion (5.7) becomes

$$\frac{d\mathbf{v}}{dt}=K\frac{\mathbf{q}-\mathbf{v}t}{t^2}. \qquad \ldots\ldots(5.10)$$

Anticipating a later result, if the test-particle moves in the presence of the substratum alone we take

$$K=-1. \qquad \ldots\ldots(5.11)$$

At any epoch t there is one observer relative to whom P is momentarily at rest; let O be this observer. Then at the epoch concerned (5.10), (5.11) reduce to

$$d\mathbf{v}/dt=-\mathbf{q}/t^2. \qquad \ldots\ldots(5.12)$$

This can be written

$$d\mathbf{v}/dt=-\tfrac{4}{3}\pi\gamma\rho\mathbf{q} \qquad \ldots\ldots(5.13)$$

where

$$\gamma=c^3t/M \qquad \ldots\ldots(5.14)$$

and ρ, M have the values given by (5.4), (5.5). The right-hand member of (5.13) is formally the same as the gravitational attraction at P due to material of density ρ within the sphere with centre O and radius OP. Thus the equation of motion can be interpreted as stating : *A free particle moves in the presence of the substratum so as always to be falling towards the apparent centre of the universe under the gravitational attraction of the material between itself and this centre.* The apparent centre means the fundamental observer in whose frame the velocity of P is momentarily zero. The attraction here means that resulting from an inverse square law with a gravitational 'constant' (5.14) proportional to the epoch. This is the expression of Milne's own first formulation of his law of gravitation (Milne 1948, §80) as it arises in our simplified presentation.

It is to be noted from (5.10) that the acceleration of P vanishes for some epoch t if, and only if, $\mathbf{v}=\mathbf{q}/t$, that is if P coincides in position and velocity with a particle of the substratum. In this case the equation of motion shows that the acceleration then vanishes for all t. (In fact the equations (5.10), (5.11) have the general solution $\mathbf{q}=(\mathbf{r}+\mathbf{s}\log t)t$ where $\mathbf{r}$, $\mathbf{s}$ are arbitrary constant vectors.) This verifies that the substratum itself moves in accordance with the equations of motion and of gravitation as discovered. In Milne's interpretation this is his justification for claiming that he did not 'neglect' gravitation in his hydrodynamical model.

Statistical system. The next stage in Milne's work was the study of what he called the 'statistical system'. This is an assembly of free particles, each moving according to Milne's analogues of (5.10). It has a space- and velocity-distribution such that (*a*) the time-variation is that produced by the actual motion of the particles, i.e. the distribution function satisfies a 'Boltzmann equation', (*b*) the system is described in the same way by all fundamental observers, i.e. it conforms to the CP.

These requirements impose another condition upon the acceleration of a particle belonging to the assembly. In the S-model this corresponds to getting

an expression for the constant K for any admissible statistical system. Using not altogether convincing arguments, Milne interpreted his result as giving, in the limiting case when the substratum alone is present, what corresponds in the S-model to the value $K=-1$. This is the source of (5.11).

With the aid of his general statistical system Milne was able to discuss his law of gravitation somewhat more fully.

Two time-scales. The study of time-keeping as the basis of kinematic relativity, which Milne developed partly in collaboration with G. J. Whitrow, is a valuable contribution to the fundamentals of all relativity theories. Its value does not depend upon its cosmological applications.

There is, however, one feature which Milne did regard as specially important for cosmology. He attached particular significance to the regraduation of clocks so as to keep τ-time instead of t-time, where

$$\tau = t_0 \log(t/t_0) + t_0, \qquad \ldots\ldots(5.15)$$

and t_0 is again some particular value of t. According to Milne's procedure, he had to define a new distance to be used in association with τ-time.

One point is that Milne's law of gravitation becomes in some respects more like the newtonian law when these new variables are used.

The general point is that, as a result of later developments of kinematic relativity which cannot be discussed here, Milne (1948, p. 224) concluded, "Statements involving Newtonian mechanics, or Newtonian gravitation, are usually made in terms of the τ-scale. Statements involving the Lorentz formulae, photons, or the electromagnetic theory of Maxwell involve the t-scale."

5.5. *Milne's Theory of Gravitation*

Milne constructed a model-universe. As we contend, this depended upon definite physical hypotheses. But these concern the behaviour of the contents of the universe as a whole. They include nothing in the nature of field-equations and, in particular, no mention of anything like gravitation. Nevertheless Milne was able to deduce certain properties, especially of the motion of a test-particle, that an observer in this universe could, and would be likely to, ascribe to 'gravitation'. In other words, an observer noticing these properties would discover the phenomenon of gravitation.

Believing, as he evidently did, that the actual universe must be a realization of his model, Milne himself claimed that he had deduced the existence of gravitation in the actual universe. He considered that he had promoted the law of gravitation from the status of an empirical result to that of a mathematical theorem. On account of the hypothetical elements to which we have called attention we have to discount this claim.

Nevertheless, Milne did show that, starting from only a few general hypotheses about the contents of the universe as a whole, it is possible to infer other properties that have something of the character of physical laws in the ordinary sense.

From this point of view it does not much matter what the particular (and perhaps peculiar!) form of Milne's law of gravitation happens to be. Actually most cosmologists would probably say that Milne chose an unsuitable model (not realizing that he had a choice). What is significant is that, using the model

he did happen to choose, Milne did demonstrate the possibility of his view that "the distribution of matter-in-motion in the universe controls and fixes the so-called laws of nature that the distribution obeys."

The following comments may now be made:

(*a*) There is nothing surprising in Milne's conclusion. We know from other theories that we can start with physical laws and construct model-universes. If desired, we could reverse the mathematics and this, in principle, is what Milne did.

However, any of the other theories referred to yields a whole class of models. Therefore we could reverse the mathematics actually employed by it only by starting with this whole class.

Now the actual universe is certainly unique. If, therefore, we have a single model, which we believe to be an acceptable model of this universe, we certainly cannot apply to it alone the reversed mathematics of one of these other theories so as to discover its physical laws. It is in fact an entirely new mathematical problem to find out what physical laws can be said to be obeyed in the model. It is this problem that Milne solved for his particular case. Even if we do not consider his model to be acceptable, we can still take it as showing *the procedure to be followed and the nature of the results to be expected.*

(*b*) From what has been said in (*a*) we obviously do not expect to discover by a procedure such as Milne's a set of physical laws as complete as those which lead to the class of models in one of the other theories. But is there any special merit in such laws as it is possible to discover by this procedure? The answer appears to be that *anything we discover in this way is necessarily in conformity with Mach's principle.*

We can illustrate this by considering the gravitational constant. This can be regarded as measuring the ratio of gravitational to inertial mass. We must suppose that such a quantity is related in a fundamental way to other properties of the physical universe. If we accept Mach's ideas it must depend upon the large-scale distribution of matter in the universe. In the other theories referred to we hope at best to discover the nature of this dependence at a late stage in the work. On the other hand, following a procedure such as Milne's we must discover this dependence as an initial feature of the law of gravitation. This is in fact shown by (5.14). (Here we are concerned only with the fact that Milne's procedure did give him an expression for γ in terms of the parameters of his model; we are not concerned with the physical validity of this particular expression.) To be in possession of a fundamental relation of this sort from the beginning of the study of gravitation may justifiably be claimed as a merit of the procedure.

(*c*) At the same time there is a serious criticism. There is *no need* for the law of gravitation as derived! For a law such as Milne's is not a complete law of gravitation. It does not enable us to write down equations of motion for any given gravitating system such as, say, the solar system. It is merely a way of interpreting properties of the model-universe derived from the postulates or hypotheses. We are not compelled to give this interpretation; if we do, we gain no new ability to make predictions.

The essence of a physical law is that it enables us to predict the results of experiments when we have a free choice, within certain limits, of the experiment to be performed. But a law derived by Milne's procedure does not enable us to predict the result of any experiment for the adequate reason that experiments are forbidden. Once the model-universe has been specified, all that an observer

is allowed to do is to observe it. He may not do experiments because, in principle, an experiment would alter the model.

That this is not a trivial criticism is seen when we examine the actual procedure itself. For Milne's procedure cannot be followed unless the CP is assumed to be *strictly* obeyed, i.e. the universe is strictly uniform. Now we know that the actual universe is uniform at most only statistically.* But Milne's procedure is not applicable to a universe that satisfies the CP only statistically. This in itself is a serious drawback. However, the point that concerns us here is that *it is precisely in order to predict the behaviour of systems which are departures from uniformity in the universe that we need physical laws. Therefore a theory depending essentially upon the assumption of uniformity can never produce the physical laws we need.*

(*d*) The conclusion seems to be that cosmology developed by a procedure such as Milne's can yield physical 'laws' which, though in themselves they add nothing to our knowledge, are more fully determined than ordinary physical laws in the sense that the constants involved are implicitly determined by the contents of the universe. On the other hand, the procedures of ordinary physics yield physical laws involving certain universal constants whose values are known only empirically. If then we are allowed to recognize the former 'laws' as constituting certain aspects of the latter, we may claim to have *a theory of the universal constants concerned*. Whether all this has much advantage over a purely extrapolatory procedure is not clear.

§6. STEADY-STATE THEORY

6.1. *General Approach*

We observe the universe from a particular position and at a particular epoch. By far the simplest attitude we can adopt is that *there is nothing special about this position or this epoch*, i.e. to assume that the universe as a whole would look the same from any other position and at any other epoch. The theory that would result from this assumption is the *steady-state theory*.

Here it is worth remarking that two general possibilities for the behaviour of the universe appeal to our minds as intuitively acceptable. But these possibilities are apparently irreconcilable. On the one hand it seems most natural to suppose that whatever we think of as the *whole* universe must always consist of the same amount of material simply because there is nowhere 'else' left for the material to go to or to come from. This makes it difficult to imagine how the universe as a whole can be changing with time. Besides, all our physical and mathematical training leads us to seek for *invariants*; we instinctively consider that the universe must have invariant properties and that the universe as a whole must almost certainly be invariant. Indeed, a semi-philosophical instinct suggests to our minds that we must in some way refer other things to the universe itself as a standard of invariance. It was preconceptions of this sort that quite naturally made it difficult to accept the theory of the expanding universe when this was first proposed.

On the other hand, our physical and mathematical ideas lead us to think of the behaviour of any system as being determined by *initial conditions*. For this

* The use of 'statistical' here is not to be confused with its use in Milne's 'statistical system'. The latter obeys the CP exactly. The customary use of 'statistical' in cosmology, as in the present paragraph, is a perhaps rather loose one intended to imply the consideration of some kind of averaged properties.

and other reasons, in the case of the universe we cannot avoid the concept of *creation*. Also we can scarcely avoid some instinctive notion that the universe is 'running down', i.e. in some way pursuing a definite 'career'. The theory of the expanding universe seemed to incorporate these ideas, and this fact contributed to its ultimate general acceptance.

Of the two, the possibility that the universe is in a steady state is certainly the more simple. Merely for this reason, we might decide to explore its consequences before proceeding to something more complicated. But if we do so we quite unexpectedly find that the requirements of the apparently conflicting view are actually met as well. We find ourselves in the rather comfortable situation that our two sets of instincts can *both* be taken as trustworthy.

6.2. *Bondi and Gold's Approach*

The steady state theory* was put forward in 1948 by Bondi and Gold (1948), but on rather more specific grounds than those indicated above.

If the writer has correctly appreciated their starting point it is this : *We must be prepared to admit that some features of the laws of physics as we know them are influenced by the large-scale properties of the universe. If these properties are changing with time, we cannot consistently apply such laws in order to study the way in which the change takes place.*

As is well known, the possibility of a dependence of local phenomena upon the large-scale distribution of matter in the universe was contemplated by Mach (1893). Bondi and Gold admit this possibility. Then, as they point out, were the large-scale distribution to vary with time or place the repetition of experiments under the same conditions would be impossible. This would call for a revision of our accepted concept of physical law, which is founded upon the assumption that such repetition is possible at will. Here is the fundamental difficulty, according to the view of these authors, in the study of a changing universe.

They do not assert *a priori* that the universe *must* be in a steady state. But they do assert that, *if* it is in such a state, then we can study it consistently in accordance with the accepted epistemology of physics. Should this study lead to conclusions in disagreement with observed properties of the actual universe, we must infer that the universe is not in a steady state. As Bondi and Gold point out, the epistemological obstacles to progress in studying the universe would then be very grave—much graver than has been recognized in formulating current theories.

Therefore Bondi and Gold emphasize the crucial importance of constructing and testing a steady-state theory. If we accept their reasoning, it is in any case the only theory that we can claim to know how to construct.

At first sight the foregoing reasoning seems to lead to a curious logical dilemma. In fact we could argue that, if the universe is in a steady state it has no physical meaning to say that local phenomena can depend upon the large-scale distribution of matter. For if the latter is always the same then its effect can never be detected by any observation or experiment. Thus we seem to have been reasoning that, if the influence of the universe as a whole is observable, so that it must vary with time or place, then we should not know how to deal with it theoretically. On the other hand, if we do know how to deal with this influence,

* The conclusions of this § 6 were reached simultaneously and independently by Hoyle (see § 9). According to the classification we are using, the Bondi–Gold approach is deductive and Hoyle's is extrapolatory ; this is the only reason for the present order of treatment.

then it must not vary, that is, it must be unobservable and therefore physically non-existent!

However, we do know that the distribution of matter in time and space in a region of the actual universe as large as, say, that containing our local cluster of galaxies is not uniform. According to the line of thought being followed, this matter could have some influence, presumably small but in principle observable, upon laboratory phenomena. What is being asserted is that, if this effect were discovered by observation and if the effect were then summed for all the matter in the universe, the result would be intelligible if, and only if (apart from the small local effects), it could be assumed independent of time and place.

It is unlikely that these rather metaphysical considerations will appeal to most physicists. But the steady-state theory turns out to be of great physical interest altogether apart from the consideration that prompted its investigation.

6.3. *Perfect Cosmological Principle*

For the reasons stated, Bondi and Gold proposed to study a universe satisfying what they call the perfect cosmological principle (PCP). This is a universe which is uniform in time as well as in space i.e. it satisfies the CP in the earlier (or 'restricted') sense and is also in a steady state. Isotropy is also assumed. The particular case of a static universe is excluded.

6.4. *Simplified Presentation*

We shall first construct an S-model satisfying the PCP. This is achieved if we require the model to be in a steady state, i.e. if we require the functions $R(t)$, $\rho(t)$, which completely determine the model, to be independent of t. Thus we make the postulates :

A. The time-keeping to be used is implied from the outset.

B. The velocity distribution is the same at all times, the fluid not being at rest.

C. The density is the same at all times.

Once again A is what we supposed in constructing the S-model. But we should notice that the postulate is here upon a somewhat different footing from what it was before. For if the universe is in a steady state it does not serve as a cosmic clock. This will be reflected in the circumstance that all formulae must be essentially independent of the origin from which t is measured. The significance of not having a cosmic time determined as in other theories has been further discussed by Bondi (1952, p. 71).

From B we have
$$V(t) = V_0 \qquad \ldots\ldots(6.1)$$
for all t, where V_0 is a non-zero constant. So here we have
$$T = 1/V_0, \qquad \ldots\ldots(6.2) \qquad Q(t) = cT \qquad \ldots\ldots(6.3)$$
for all t, where T is a constant, as before, which now has no reference to any particular epoch t_0.

Using (4.5) we now obtain
$$R(t) = e^{t/T}. \qquad \ldots\ldots(6.4)$$
Here we have taken $R(0) = 1$; but a change of origin of t only introduces a constant factor in R, and so does not affect the velocity distribution derived from (4.3).

It might have been thought that, if the universe is in a steady state for one method of time-keeping, it must be in such a state for any other method, and

consequently that $R(t)$ ought to be independent of the time-keeping. This is false since velocities are concerned. If, as we are supposing, the fluid-velocity at a given position in an observer's frame is independent of the epoch for one method of time-keeping, this will not be true for any essentially different method.

This does, however, suggest a possibly better way of regarding the postulates. If any S-model has $R(t)$ monotonic increasing, then we can transform t so that (6.4) holds good. Then we could postulate that it is this time that plays the part of newtonian time and is also that for which the speed of light is constant. The suggestion would be that the influence of the large-scale distribution of matter in the universe upon local phenomena is such as to conceal, or make physically meaningless, any departure from large-scale uniformity over time and space.

It is interesting to note that this would convert Bondi and Gold's PCP into a *principle of impotence* of the sort remarked upon by Whittaker (1942). It could be formulated: *It is impossible to detect any departure from the perfect cosmological principle.*

Even if acceptable, this could be only a provisional formulation. As it stands, it could apply only to a large-scale departure from the PCP. A more exact statement would be expected, by allowing for the necessary detectability of local departures, to imply also a law of gravitation.

Any such alternative standpoint would have no effect upon the inferred properties of the model.

6.5. *Creation*

If ρ_0 is the constant density required by C, then we have from (4.11), (6.3)

$$M=\tfrac{4}{3}\pi c^3\rho_0 T^3, \qquad \ldots\ldots(6.5)$$

which is another constant.

Now at epoch t let $\mathcal{V}(t)$ be the volume within any closed surface moving with the fluid. As before, the linear dimensions of $\mathcal{V}(t)$ are proportional to $R(t)$, so that

$$\mathcal{V}'/\mathcal{V}=3R'/R. \qquad \ldots\ldots(6.6)$$

Since the density, not the total mass, within $\mathcal{V}(t)$ is assumed to be constant, it follows that material must be 'created' in $\mathcal{V}(t)$ at the rate $\rho_0\mathcal{V}'(t)$. Hence, using (6.1), (6.6), *the required rate of creation is* μ *per unit volume per unit time where*

$$\mu=\rho_0\mathcal{V}'/\mathcal{V}=3\rho_0/T. \qquad \ldots\ldots(6.7)$$

Moreover, in order to preserve the density, this 'created' material must be in possession of the velocity of the material already present and to which it is added. That is, the 'created' material is at rest relative to the fundamental observer at each point where it makes its appearance.

The presentation here is arranged so as to make it clear, as emphasized by Bondi (1952, p. 6), that we do not need any greater number of postulates than there are in, say, Milne's theory. In particular, both theories require a postulate C; it is merely different in the two cases, and there can be no logical objection to either. What we have now expressed is the obvious inference that the postulate of constant density in a non-static model requires that material of the sort measured by this density should be added to the system at a certain rate. This appearing in the system of material not previously there we term 'creation'.

At epoch $t-\alpha$ the value of $\mathcal{V}(t)$ was $\mathcal{V}(t-\alpha)$. Hence, of the amount of material $\rho_0\mathcal{V}(t)$ present at epoch t, an amount $\rho_0\mathcal{V}(t-\alpha)$ was already present at

epoch $t-\alpha$, since there has been no destruction of material in the interval. Therefore the fraction of material having age exceeding α is

$$\frac{\mathcal{V}(t-\alpha)}{\mathcal{V}(t)} = \left[\frac{R(t-\alpha)}{R(t)}\right]^3 = e^{-3\alpha/T}. \qquad \ldots\ldots(6.8)$$

Consequently the fraction having ages α to $\alpha+d\alpha$ is

$$\frac{3}{T}e^{-3\alpha/T}\,d\alpha. \qquad \ldots\ldots(6.9)$$

It follows that

$$\textit{the average age of material in any region} = \tfrac{1}{3}T. \qquad \ldots\ldots(6.10)$$

The foregoing results are collected in the table.

6.6. *Space–Time*

We have remarked that a space–time of the form (4.17) is acceptable in all the theories we are discussing. In order to conform to the PCP of the present theory, the metric must transform into itself under a change of origin of t. As Bondi (1952, p. 146) shows, this requires

$$k=0, \qquad R(t)=e^{t/T}. \qquad \ldots\ldots(6.11)$$

Thus the form of R agrees with that of our simplified presentation. Moreover, the space of the world-map is euclidean. Indeed, all the properties of our S-model as obtained are formally identical with those got by Bondi and Gold. The S-model is deficient only in some details of the calculation of the world-picture from the world-map.

The model will be discussed further after it has been derived in another way.

§ 7. NEWTONIAN UNIVERSES

7.1. *Extrapolatory Theory*

In an extrapolatory theory we take it that we know the laws of physics for some neighbourhood of the universe. We might then ask how large this neighbourhood may be taken to be, or, as Bondi (1952, p. 3) puts it, "What is the largest set of objects to which our physical laws can be applied consistently and successfully?" We might, indeed, find it reasonable to *define* this set to be the physical universe. The emphasis would thus be upon the laws of physics, and the universe would be the totality of whatever obeys these laws.

However, it is difficult to proceed strictly along these lines. What we are calling extrapolatory theories follow in practice a somewhat more mixed procedure. They enquire whether a universe which satisfies the CP can be regarded as the totality of objects obeying the known laws of physics. More precisely, they postulate that the universe obeys the CP and that the contents obey the known laws of physics and investigate whether a model without inconsistencies is thereby determined.

It is important to remark that the CP can here be regarded as playing a less fundamental part than in the deductive theories. It is legitimate, if we wish, to regard it as no more than a tentative condition which should lead to the simplest type of solution of our problem, if in fact it can be satisfied.

7.2. *Newtonian Universe*

A newtonian universe will be defined as *an S-model in which the material moves in accordance with newtonian mechanics under its newtonian self-gravitation.*

Since newtonian space and time are used in the S-model, this now becomes something *per se*; it is not here used as a simplified presentation of anything else.

There may be a stress in the fluid. From the assumption of isotropy this can only be a (positive or negative) pressure. By §4.3 (*a*) this is a function $p(t)$ of t only. Hence the pressure gradient is zero and the pressure does not contribute to the force-term in the equation of motion.

We take $\rho(t)$ to be the density of *inertial mass*. Since we are using newtonian theory, we assume the conservation of such mass.

In newtonian theory an observer must know what he is to take as a newtonian frame of reference. Now in the S-model any fundamental observer O regards the whole universe and its motion as spherically symmetric about himself in his F-frame. Hence he cannot do otherwise than treat the F-frame as one in which the newtonian laws are valid, provided he can use these laws at all.

We might say that, if we allow observer O to do this, we ought not to allow another observer to do likewise because his frame would in general be accelerated relative to O's. However, we shall find that this makes no difference. This point has been more fully discussed by Bondi (1952, p. 78, §9.3).

The laws concerned are expressed by the equation of continuity and the equation of motion. The former gives, as in (5.3),

$$\rho = \rho_0 / R^3. \qquad \ldots\ldots(7.1)$$

From (4.7), (4.8) the acceleration of a particle of the fluid at position $\mathbf{q}$ is

$$\frac{d^2\mathbf{q}}{dt^2} = \frac{R''}{R}\mathbf{q}. \qquad \ldots\ldots(7.2)$$

If there were a distribution of matter of gravitational mass-density $\sigma(t)$ through a sphere with centre O and any finite radius greater than q, the intensity of newtonian gravitation at q would be

$$-\tfrac{4}{3}\pi\gamma\sigma\frac{q^2}{q^2}\left(\frac{\mathbf{q}}{q}\right) = -\tfrac{4}{3}\pi\gamma\sigma\mathbf{q} \qquad \ldots\ldots(7.3)$$

where γ is the newtonian gravitational constant. We interpret the newtonian law as continuing to give this result in the case where the distribution extends to infinity provided the distribution continues to exhibit spherical symmetry about O. Also standard newtonian theory identifies gravitational and inertial mass so that

$$\sigma = \rho. \qquad \ldots\ldots(7.4)$$

Combining (7.2), (7.3), (7.4) and remembering that the stress produces no resultant force, the equation of motion becomes

$$R'' = -\tfrac{4}{3}\pi\gamma\rho_0/R^2. \qquad \ldots\ldots(7.5)$$

This integrates to give

$$R'^2 + kc^2 = \tfrac{8}{3}\pi\gamma\rho_0/R \qquad \ldots\ldots(7.6)$$

where the integration constant kc^2 is written thus for future convenience.

Rearranging (7.1), (7.5), (7.6) we obtain

$$3\frac{R'^2 + kc^2}{R^2} = 8\pi\gamma\rho \qquad \ldots\ldots(7.7)$$

$$2\frac{R''}{R} + \frac{R'^2 + kc^2}{R^2} = 0. \qquad \ldots\ldots(7.8)$$

It is to be noticed that the vector **q** cancels out when we combine (7.2), (7.3). Thus we are left finally with only the two equations (7.7), (7.8) for the functions $R(t)$, $\rho(t)$. These equations are independent of the choice of the observer O. Therefore any functions $R(t)$, $\rho(t)$ satisfying these equations give an S-model that meets all our requirements. This verifies the consistency of these requirements. Therefore *newtonian model universes as defined do exist as mathematical possibilities.*

7.3. *Remarks*

Significance of newtonian models. Until comparatively recently it was believed that an infinite newtonian universe, and consequently one satisfying the CP, is an impossibility. It was while this view was still accepted that relativistic cosmology was developed with apparent success. Therefore it came to be thought that the large-scale behaviour of the universe cannot be understood apart from relativity theory. However, without here going into all the arguments, it now seems fair to assert that there is nothing essential in the relativistic treatment that cannot be expressed in newtonian terms.

It is easy to see why the infinite newtonian universe was considered impossible. For it was formerly taken for granted that a uniform distribution of matter throughout euclidean space must be static. Now a static model has $V=0$ giving $R'=0$ in (4.3), and we see from (7.7), (7.8) that this leads to $\rho=0$. *Hence a non-empty static newtonian universe is impossible.* But we now see that we can proceed quite consistently to get *non-static* models.

Thus the physically simplest theory of the universe, the newtonian theory, does not allow a static universe but does allow a non-static universe. At the same time the simple interpretation of the observed red-shifts in the spectra of the galaxies is that the actual universe is non-static, in fact expanding. The combination of these results furnishes perhaps the most convincing argument for accepting the concept of the expanding universe.

Another reason why it is still sometimes thought that the relativistic treatment is indispensable is the contention that it alone admits the notion of expanding space. But the relativistic notion of the universe expanding *with* space and the classical notion of its expanding *in* space are only different mathematical ways of describing essentially the same observable physical phenomenon.

Constant k. From (7.6) it can be verified that there is one, and only one, newtonian universe for each pair of values of ρ_0, k. Physically, this means that any uniform distribution of matter, if set in motion with velocity proportional to distance from any particle of itself, will proceed to move in accordance with the assumed laws, so that the motion continues to be of this character. The factor of proportionality at any epoch is R'/R where R is the solution of (7.6) (for which we can take $R=1$ at the epoch of 'projection' in order to dispense with a superfluous constant). As can be seen from (7.6), the meaning of k is that for $k<0$, $=0$, >0 any observer describes any particle as having at any epoch a velocity greater than, equal to, or less than the velocity of escape from the material between himself and that particle. As we should then expect, if the system is initially expanding, it will continue to expand in the first two cases, while the motion will be oscillatory in the third case.

Problem of uniqueness. A feature calling for remark, and one which is common to all extrapolatory treatments, is that *the theory does allow more than one*

model (in the present case a double infinity of models). In accordance with orthodox physical practice we should proceed to examine whether any one of the theoretical models gives results in satisfactory agreement with observation. Also we might find ancillary reasons for seeking to identify one of the models with the actual universe. For instance, since the case $k=0$ is that in which, in the smoothed-out universe, every particle has just the velocity of escape from the rest of the material, Hoyle (1948) has argued for this being the case in which perturbations of the material would produce condensation into galaxies of the sort actually observed.

There seems to be no reason why a model selected in this way, if the agreement with observation is sufficiently good, should not be regarded as giving physical understanding of the actual universe. We could accept it as giving an astrophysical theory of the large-scale behaviour of the astronomical universe on the same footing as any other item of accepted astrophysical theory.

Nevertheless we should not regard the understanding so obtained as really fundamental. The conception of the actual universe being selected from an infinity of universes that could exist, by the accident of a particular set of initial conditions, is unsatisfying. In fact most physicists probably hold the view that *we cannot attach a meaning to the possible existence of any universe other than the actual universe.*

7.4. *Generalized Newtonian Models*

We can now make a newtonian adaptation of two results of relativity theory: (*a*) the equivalence of mass and energy, (*b*) the difference between gravitational and inertial mass for material under stress (McCrea 1951 a). Also we can allow for the possibility of 'cosmical repulsion' (McCrea and Milne 1934), which is in fact the only other type of force, other than newtonian gravitation, that does not violate the consistency of the equations. We then get in place of (7.7), (7.8)

$$3\frac{R'^2+kc^2}{R^2}-c^2\lambda=8\pi\gamma\rho, \qquad \ldots\ldots(7.9)$$

$$2\frac{R''}{R}+\frac{R'^2+kc^2}{R^2}-c^2\lambda=-\frac{8\pi\gamma}{c^2}p \qquad \ldots\ldots(7.10)$$

where λ is Einstein's cosmical constant.

The only difference from (7.7), (7.8) is the appearance of λ in the left-hand members and the appearance of p on the right in the second equation. The latter feature is a novelty from the newtonian standpoint since, in any strictly newtonian equation of motion, the absolute value of the stress never appears; only the gradient of the stress is ever involved.

If the nature of the material is specified by a given relation between ρ, p, and if λ is a given constant, then there exists a double infinity of solutions as before. Those for which $p=0$ have usually been regarded as of chief interest and have often been described (Tolman 1934, Heckmann 1942, Bondi 1952).

The derivation of (7.9), (7.10) is not given here since the same equations will arise later in relativistic cosmology. However, we proceed to give several special solutions here rather than in the latter connection merely because we want to discuss them in newtonian terms. The mathematical results are formally the

same as in the relativistic cases and so we use the same names for the cases as they have acquired in relativistic cosmology.

7.5. *Particular Models*

(*a*) ***Einstein universe.*** Here we take $p=0$, $\lambda \neq 0$ and look for a static solution, i.e. we must have $R=1$, as we see from (4.3), (4.4). We then get from (7.10), (7.9) respectively

$$k=\lambda \qquad \ldots\ldots(7.11)$$

$$\rho=\rho_0=\lambda c^2/4\pi\gamma. \qquad \ldots\ldots(7.12)$$

The meaning of (7.12) is that, for this value of ρ, ***gravitational attraction is exactly balanced by cosmical repulsion*** at all distances from any observer. This is seen by going back to (7.5) which reads, with the cosmical term included and using (7.12),

$$R''/R=-\tfrac{4}{3}\pi\gamma\rho_0/R^3+\tfrac{1}{3}\lambda c^2=\tfrac{1}{3}\lambda c^2(1-1/R^3). \qquad \ldots\ldots(7.13)$$

We see from (7.13) that, if the system is released with $R=1$, then R'' remains zero and the system remains in equilibrium. If it is released with $R<1$, then $R''<0$, so that R proceeds to decrease still further; similarly, if it is released with $R>1$ then R proceeds to increase. Thus ***the equilibrium state is unstable*** (Eddington 1930).

(*b*) ***Eddington–Lemaître universe.*** This is the model into which the Einstein universe passes when the equilibrium is disturbed, so that R proceeds to increase. More accurately, it is the expanding model which has the Einstein universe as the limit for $t\to-\infty$. As follows from (7.13), for large positive t it has

$$R\sim\exp[(\tfrac{1}{3}\lambda)^{1/2}ct]. \qquad \ldots\ldots(7.14)$$

(*c*) ***De Sitter universe.*** If we take $\rho=0$, $p=0$, $\lambda>0$ we get from (7.9), (7.10)

$$R=\exp[(\tfrac{1}{3}\lambda)^{1/2}ct] \qquad \ldots\ldots(7.15)$$

if we take the positive sign of $\lambda^{1/2}$.

We can interpret this case as that in which the density is so small that ***the only significant force is cosmical repulsion.*** It may be said that historically it was de Sitter's (1917) discovery of this case that predicted the phenomenon of the expanding universe. When the phenomenon was established empirically, more than ten years later, this was naturally hailed as a triumph for the theory of general relativity employed by de Sitter. This prediction helped to implant the idea that the expansion is a manifestation of a physical phenomenon that can only be understood through relativity theory. But the situation is doubly odd. In the first place, all that de Sitter did from the physical standpoint was to consider a system moving under cosmical repulsion alone, and so it is not surprising that he got an expanding universe. Also we now see that, if we do want to postulate cosmical repulsion, we can do so just as well in a classical as in a relativistic treatment. In the second place, even in relativistic cosmology, the idea of this cosmical repulsion has now been generally abandoned (but see Lemaître 1950).

Another interest of de Sitter's universe is that it is the asymptotic form of the Eddington–Lemaître universe for large t, as is seen from (7.14), (7.15), or indeed of any model in which cosmical repulsion exceeds gravitational attraction.

We now note from (4.3), (4.9), (4.10), (4.11) with (7.15) that

$$V=(\tfrac{1}{3}\lambda)^{1/2}c=1/T, \qquad \ldots\ldots(7.16)$$

$$Q=(\tfrac{1}{3}\lambda)^{-1/2}. \qquad \ldots\ldots(7.17)$$

Also, if we allow a small density satisfying (7.1) but too small to have any gravitational effect, we get

$$\rho = \rho_0 \exp[-(3\lambda)^{1/2}ct], \qquad \ldots\ldots(7.18)$$

and so from (4.11)

$$M = \tfrac{4}{3}\pi\rho_0(\tfrac{1}{3}\lambda)^{-3/2}\exp[-(3\lambda)^{1/2}ct]. \qquad \ldots\ldots(7.19)$$

It should be remembered that any metric of the form (4.17) with R an exponential function of t is commonly called a de Sitter *metric*, whether or not it is associated with the de Sitter *universe*.

(*d*) *Einstein–de Sitter universe.* This is the simplest non-empty model and is got by taking $k=0$ in (7.7), (7.8) or $k=0$, $p=0$ in (7.9), (7.10). The solution having $R'>0$ and referred to a suitable origin of t is

$$R = (t/t_0)^{2/3}, \qquad \ldots\ldots(7.20)$$

$$\rho = (6\pi\gamma t^2)^{-1}. \qquad \ldots\ldots(7.21)$$

From (4.9), (4.10), (4.11) we then get

$$T = \tfrac{3}{2}t_0, \qquad \ldots\ldots(7.22)$$

$$Q = \tfrac{3}{2}ct, \qquad \ldots\ldots(7.23)$$

$$M = \tfrac{3}{4}c^3\gamma^{-1}t. \qquad \ldots\ldots(7.24)$$

(*e*) *Milne universe.* If we formally insert (5.6) in (7.9), (7.10) with $\lambda=0$, noting that we have not yet established a connection between the two uses of the constant k, we see that the latter equations are satisfied if, and only if, $\rho=0$, $p=0$. Thus Milne's universe is kinematically equivalent to an empty universe of the present category. Here there is no cosmical repulsion, and the physical interpretation is simply that, from the present standpoint, *the model represents non-gravitating matter dispersing in space*; hence the velocity of any particle is uniform, in agreement with Milne.

If again we allow a density satisfying (7.1), but too small to have any gravitational effect, we recover (5.3), (5.4), (5.5).

(*f*) *Steady-state model.* We take $\lambda=0$ and enquire whether the Bondi–Gold–Hoyle steady-state model provides a solution of (7.9), (7.10). Remembering that this model has $k=0$, $R=e^{t/T}$ (though again we have not yet connected the uses of k), we find that the equations are satisfied by

$$8\pi\gamma\rho = \frac{3}{T^2} \quad \ldots\ldots(7.25), \qquad 8\pi\gamma\frac{p}{c^2} = -\frac{3}{T^2}. \quad \ldots\ldots(7.26)$$

Bondi and Gold do not use any field equations and, so far as their theory has been developed, their values of ρ, T are not related. We now see that their model does satisfy the generalized newtonian equations if ρ, T are related by (7.25) and if the fluid (or 'medium' as it might now be more suitably called) is under the negative pressure given by (7.26).

We consider now the physical interpretation of what we have found, consideration of its possible significance being deferred. We note first that the pressure is not directly observable through any local mechanical effects since only a non-uniform pressure can produce such effects.

Consider then any sphere of fixed radius centred on the observer O. According to our previous discussion of the model, O asserts that material is continually receding across the surface of the sphere and being replaced by material created throughout the interior. But we can now add that, by virtue of the negative

Model Universes

Model		M.H.	B–G–H	E.S.	E–L	de S	E–de S	Num. val.
Recession ' constant ' $\propto$ vel. at particular *distance*	$V(t)$	t^{-1}	T^{-1}	0	$\sim(\frac{1}{3}\lambda)^{1/2}c$	$(\frac{1}{3}\lambda)^{1/2}c$	$(\frac{2}{3}t)^{-1}$	$1{\cdot}8\times10^{-17}$ sec^{-1}
Distance function $\propto$ dist. of particular *galaxy*	$R(t)$	t	$\exp(T^{-1}t)$	1	$\sim\exp[(\frac{1}{3}\lambda)^{1/2}ct]$	$\exp[(\frac{1}{3}\lambda)^{1/2}ct]$	$t^{2/3}$	
Cosmical constant	λ	0	0	λ	λ	λ	0	0
Age of universe	t_0	T	∞*	∞	∞	∞	$\frac{2}{3}T$	
Radius of observable universe	$Q(t)$	ct	cT	∞	$cT=c(R/R')$	$(\frac{1}{3}\lambda^{1/2})^{-1}$	$\frac{2}{3}ct$	$1{\cdot}7\times10^{27}$ cm
Density	$\rho(t)$	$\rho_0T^2t^{-3}$	$\rho_0=(\frac{8}{3}\pi\gamma T^2)^{-1}$	$\rho_0=\lambda c^2(4\pi\gamma)^{-1}$	$\rho_0(R_0/R)^3$	0	$(6\pi\gamma t^2)^{-1}$	$5{\cdot}8\times10^{-28}$ g cm^{-3}
Pressure	$p(t)$	0	0 or $-\rho_0c^2$	0	0	0	0	
Mass of observable universe	$M(t)$	$\frac{4}{3}\pi c^3\rho_0T^3$	$\frac{4}{3}\pi c^3\rho_0T^3=\frac{1}{2}\gamma^{-1}c^3T$	∞	$\frac{4}{3}\pi c^3\rho_0(R_0/R')^3$	0	$\frac{2}{7}\gamma^{-1}c^3t$	$1{\cdot}1\times10^{55}$ g
Rate of creation	μ	0	$3\rho_0T^{-1}=(\frac{8}{3}\pi\gamma T^2)^{-1}$	0	0	0	0	$3{\cdot}1\times10^{-44}$ g cm^{-3} sec^{-1}
Independent parameters		T, ρ_0	B–G : T, ρ_0 H : T	ρ_0	ρ_0	λ	T	

M.H., Milne hydrodynamical system; B–G–H, Bondi–Gold–Hoyle steady-state universe; E.S., Einstein static; E–L, Eddington–Lemaître; de S, de Sitter; E–de S, Einstein–de Sitter; Num. val., numerical values for steady-state universe.

* Mean age of contents $\frac{1}{3}T$.

pressure in the moving medium, the medium outside the sphere is doing work upon the medium inside. The meaning of our equation of continuity in this case is that this work is exactly equivalent to the energy-equivalent of the material coming out. *Thus the net flow of energy across the surface is zero.* We can think of the medium as expanding under such a tension that it continually replenishes itself by the work it does upon itself. Therefore the process is energetically self-supporting. The only novel feature in the creation process is the supposition that the energy represented by the work done makes itself apparent in the form of, or is continually being converted into, matter of the same sort as the matter already present.

Furthermore, the relation between gravitational and inertial mass employed here is such that the negative pressure of the amount calculated makes the gravitational mass negative. This in turn makes the gravitational force negative, and of the amount necessary to produce the outward acceleration maintaining the steady velocity-distribution. (It is seen that, if a particle of the fluid recedes from O in such a way that its velocity is always a constant multiple of the distance, it is necessarily accelerated away from O.) Thus the force producing the motion is accounted for. The system is therefore completely self-supporting.

§8. RELATIVISTIC UNIVERSES

8.1. *Comparison with Newtonian Theory*

Relativistic universes are got by the procedure of the previous section when we employ the concepts of general relativity theory in place of those of classical theory. It is only a convention to reserve the term 'relativistic' for these since some form of 'relativity' is necessarily used in every theory.

As already stated, general relativity requires a model which satisfies the CP to have a space–time of the form given by (4.17). But now general relativity does away with the classical separation between space and time on the one hand and matter on the other. It treats these as properties of the single entity 'space–time'. In general relativity this entity is assigned a much more comprehensive significance than what is called space–time in classical theory or in special relativity theory. Given the metric of any space–time, Einstein's field relations, usually written as

$$G_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} G + \lambda g_{\mu\nu} = -\kappa T_{\mu\nu}, \qquad (\kappa = 8\pi\gamma/c^2), \qquad \ldots\ldots (8.1)$$

tell us what material system is specified by it. The left-hand member of (8.1) is determined entirely by the metric, and the tensor on the right is the stress–momentum–energy tensor of the material. We say that the system so found is moving in accordance with Einstein's law of gravitation.

Applied to the metric (4.17), it is found that the *relations* (8.1) *give precisely the equations* (7.9), (7.10) when the constant k in these equations is the constant k in the metric.

This means that, according to general relativity, the system described by the metric for any given $k, R(t)$ is a distribution of matter of density given by (7.9) with a stress which reduces to the pressure given by (7.10). Also the momentum components of the material tensor are found to be zero, which denotes that the material is at rest in the r, θ, ϕ coordinate system: this is consistent with the basic interpretation of the metric given on p. 331 before any field-relations were used.

Consider first the cases for which $p=0, \lambda=0$. ***Then the relativistic equations are identical with the simple newtonian equations*** (7.7), (7.8). This (at first sight surprising) result is highly instructive. For it shows that the relativistic treatment deals with exactly the same physical situation as does the newtonian. If we do adopt a relativistic cosmology we can therefore still describe it physically in newtonian language, and actually derive the main quantitative relations by newtonian calculations.

The explanation of the identity of the results of the two treatments is briefly that, as is known, the two laws of gravitation are equivalent for 'weak' fields and that this equivalence extends to the present case because (7.7), (7.8) can be regarded as expressing conditions in the neighbourhood of a point in the field. In particular we can in both treatments interpret k in the same way in terms of the escape-velocity. However, the newtonian scheme can always accommodate the different types of motion in its unique euclidean space, whereas the relativistic scheme can accommodate them only with the assistance of an appropriate curvature of space. Thus the two treatments differ somewhat in the manner in which they piece together what is happening in various neighbourhoods.

There is no doubt that ***the relativistic treatment is superior to the classical one.*** As explained, it includes a satisfactory treatment of light-propagation which is lacking in the classical case, though this has nothing to do with the field-relations. In regard to the latter, in order to get a consistent classical scheme the treatment of newtonian frames and infinite distributions had to be somewhat heuristic. But the relativistic treatment proceeds automatically. Nevertheless, the newtonian treatment is important for understanding the physical situations that arise.

Coming to the cases for which $p\neq0, \lambda\neq0$ we meet no fresh surprises because the extension of the newtonian results was obtained by borrowing certain results from relativity theory. Here the persistence of the identity of the equations in the two treatments is a verification that we gave the borrowed results their correct physical interpretation.

8.2. *Observable Relations*

The only effective mathematical difference between a relativistic model and the corresponding newtonian one is, as we have just noted, that in the former, but not in the latter, the geometry of the model depends upon the value of the constant k. In considering the interpretation of this difference we encounter some important problems which it is convenient to deal with at this stage.

While newtonian theory itself provides no treatment of light-propagation, it certainly allows us to assume that, relative to a particular observer O the speed of light is a constant c independent of the light-source and that the classical formula for the doppler effect is valid. This would enable us to compute the world-picture for O.

The relativistic theory does include a theory of light-propagation. A proper application of this theory to any model with its appropriate geometry suffices for the computation of the world-picture in this case.

The world-picture is the aggregate of what are called 'observable relations', such as the relation between spectral red-shift and apparent magnitude for the galaxies, the relation between numbers of galaxies and apparent magnitudes, etc. Much investigation has been devoted to these (Tolman 1934, Hubble 1936, 1937, McCrea 1935, McVittie 1939). They provide the only means for testing the theories by observation of the actual universe.

It seems legitimate to assert in a general way that the world-pictures got in the ways described for a newtonian model and the corresponding relativistic model are not very different even when there is a considerable difference between the world-maps.

As an example we may consider the Einstein static universe. The newtonian world-map is an infinite euclidean space uniformly filled with material at rest. An observer's world-picture must show him an infinite universe. The relativistic world-map, on the other hand, is a finite hypersphere uniformly filled with material and containing therefore a finite amount of this material. But an observer in this model sees any particle of the material an infinite number of times because light can travel round the universe an infinite number of times. Thus the world-picture again shows an infinite universe. (The device of regarding the geometry as elliptic instead of spherical can be made to alter the conclusion only at the price of what amounts to an additional physical assumption.)

The importance of this example is in showing that the explanation of our not seeing an infinite universe, in the sense that we do see celestial objects against a relatively *dark* background in the night sky, is not to be found in the curvature of space. Any static model-universe hitherto proposed fails to explain this elementary observation (Bondi 1952, chap. III).

On the other hand, if $V(t)>0$ for any newtonian model, so that it can be described as an *expanding* universe, then there exists the distance we have called $Q(t)$. This is such that the material at this distance from an observer O is receding from him with speed c. According to the classical theory of radiation, a light-source moving with this material has zero intensity for O and so cannot be seen by O. *A fortiori*, material beyond $Q(t)$ cannot be seen by O. Hence O can see only a finite universe. If the material is parcelled into discrete luminous objects, these must therefore show up bright against the background of the sky. The observed state of affairs is thereby accounted for.

The finiteness of the observable universe does not imply that it must appear to O to be cut off at a finite distance. We have noticed that a light-source at the 'boundary' distance $Q(t)$ has vanishing luminosity for O. Thus it is 'at infinity' according to the astronomer's usual way of judging distance by apparent magnitude. Therefore the appearance is that of a universe tailing-off to infinity but containing only a finite total amount of matter.

In the corresponding relativistic model all this is handled in a more tidy fashion but the results are similar.

The world-pictures according to the two theories are not in general identical. However, it can be said that if the density and velocities are of the order of those in the region of the actual universe accessible to present means of observation the observable relations for corresponding models in the two theories differ only in second-order effects. The differences are probably all within the limits of uncertainty of present observations and calculations. The conclusion is, in brief, that '*the curvature of space*' *is in principle observable but that even its sign cannot be definitely inferred from existing data.*

The uncertainties in the calculations arise because it is necessary to make some assumptions about what is being observed in a model. In calculating observable relations it is usual to suppose the material to be replaced, wholly or in part, by a distribution of discrete bodies such that the density of matter in the model is unaltered. These bodies are then supposed to be self-luminous

and certain assumptions are made about their luminosities and colours and the dependence of these upon epoch. These assumptions are intended to make the bodies represent the galaxies in the actual universe. The final comparison with observation has to test such assumptions as well as the general characteristics of the model. It will be seen that the situation becomes complicated and uncertain. The only case in which appreciable simplification occurs is that of the steady-state universe, there being in that case no dependence upon epoch. The reader may refer to the calculations for the steady-state model (Bondi and Gold 1948, Bondi 1952, §12.4) both as an example of these problems and for the special interest of this particular case.

§9. HOYLE'S THEORY

About the same time that Bondi and Gold produced their steady-state theory Hoyle (1948, 1949a, b) produced a different theory leading to effectively the same model. The two theories are, in fact, 'deductive' and 'extrapolatory' approaches to almost the same results.

Einstein's field-relations (8.1) require the conservation of matter in a certain generalized sense. Therefore it appears that these relations cannot take account of the creation of matter. But relativistic cosmologies, which employ Einstein's relations, make the present state of the universe depend upon the circumstances of its 'creation' in some singular state at a finite past epoch. That is to say, these cosmologies depend upon phenomena that they are inadequate to discuss. This seems to be the trend of Hoyle's argument showing that a logically satisfactory relativistic cosmology ought to employ field-relations that can take explicit account of a creation process.

Hoyle apparently regards his work as tentative in its mathematical formulation. So he looks for the least modification in Einstein's formulae that causes them to admit possibilities of the required sort. Since Einstein's conservation laws result from the fact that the left-hand member of (8.1) has identically vanishing divergence, Hoyle proposes to introduce some additional symmetric tensor of non-vanishing divergence. He drops the cosmical terms in λ, since these represent a different kind of extension of Einstein's original form of his relations from the one now desired.

Hoyle succeeds in finding a new tensor of the required sort, using derivatives of a vector-field defined at all events in space-time. His procedure is of great interest and repays close study. However, we need not describe it further here because, so far as we need consider its consequences in the model he derives, *these are identical with those of* §7.5(*f*). The only difference concerns a point of interpretation; what we identified as negative pressure Hoyle identifies with the only non-vanishing part of his new tensor in this application, and he takes the pressure itself to be zero. Apart from interpreting his results in this way, *his model is therefore identical with the Bondi–Gold steady-state model with the addition of the relation* (7.25) *between* ρ, T. However, since it is reached by an extrapolatory procedure, Hoyle's model is not a unique solution of his equations. It is the only steady-state solution, and he gives strong reasons for selecting it.

§10. PHYSICAL DISCUSSION

In this section we discuss the general physical features of the actual universe that may be inferred from its comparison with the various theoretical models.

10.1 *Preliminary Results*

(*a*) *Cosmological principle.* Astronomers may tentatively assume that space is euclidean and that the distances they measure are euclidean distances and, further, that the galaxies they observe in one region are intrinsically like those they observe in another. They can then ask whether the galaxies are uniformly distributed (statistically) throughout the volume of space accessible to observation, allowing for plausible limits of uncertainty in these assumptions. This certainly appears to be the case (Hubble 1936, 1937). So we take it that *there is prima facie evidence that the actual universe conforms to the CP* and that therefore it is reasonable to compare it with models conforming to the CP.

(*b*) *Recession of the galaxies.* Except for Einstein's static model all models yield a 'Hubble's law', i.e. a velocity of approach or recession proportional to distance. The actual universe obeys Hubble's law in the form stated (3.1), and we provisionally accept the interpretation of the observations implied. So we confine attention to expanding models or models in expanding phases. We take the parameter t_0 to denote the epoch of observation. Then the constant we have called T is the reciprocal of Hubble's constant, so that, approximately, $T = 2 \times 10^9$ years.

(*c*) *Density.* In the actual universe the energy in the form of radiation is negligible for present purposes compared with that in the form of matter. So the density ρ in a model is to be compared with the density of matter in the actual universe.

(*d*) *Pressure.* The pressure p in a model is usually taken to be compared with the kinetic pressure due to the peculiar motions of the galaxies, averages being taken over suitable regions. Since these motions are small compared with c, the resulting value of p/c^2 is very small compared with ρ. So the pressure term in equations like (7.10) is usually put equal to zero.

However, as we have said, the pressure itself does not appear in the classical equations. Also there is no necessity for the classical evaluation to fix the zero from which the relativistic value is to be measured. In other words, relativity theory may need to take account of a 'zero-point pressure' or 'rest-pressure'. So the evaluation $p = 0$ need not be treated as definitive.

(*e*) *Cosmical constant.* After the inadequacy of Einstein's static model was demonstrated, the chief argument for including the λ-terms in the field-equations has ceased to exist. Einstein (1951, p. 106) appears to be of this opinion. If the sole consideration is to find the relativistic model that best fits a given set of observational results, it may be argued that the correct procedure is not to assume λ to be zero but to treat it as one of the parameters to be evaluated. This has been the procedure commonly followed in the past. But most relativists seem now to hold that there is no general physical argument for the existence of the cosmical repulsion denoted by the λ-terms and therefore no justification for a non-zero λ.

The device of a 'zero-point' stress is in itself not unlike the use of a non-zero λ. But it can be regarded as mathematically the simplest way of allowing for a creation process and it can be held that there are general physical arguments for the existence of this phenomenon.

(*f*) *Results.* Taking the values of T, p, λ indicated, the table gives the resulting values of the other parameters for various models. -

10.2. *Derived Results*

(*a*) *Density.* Models that do relate the density to T give values of the order of that given in particular by the Hoyle model, i.e. $\frac{3}{8}(\pi\gamma T^2)^{-1} \simeq 6 \times 10^{-28}\,\mathrm{g\,cm^{-3}}$. The fact that, if they give any relation, it is of this type is not surprising since it is the only expression likely to arise having the required dimensions. Nevertheless, we may find it remarkable that observations of the radiation from distant galaxies should enable us to draw any conclusion at all about the density of matter in the universe. The fact that the inferred value is quite probably of the correct order of magnitude is good evidence for the correctness of the interpretation of the observations, for the fundamental importance of the phenomena so revealed, and for the general validity of the theories.

It is usually estimated that the galaxies account for a mean density of about $10^{-30}\,\mathrm{g\,cm^{-3}}$, though this may not make enough allowance for the non-luminous matter they contain. But there is a growing recognition that there may be much more matter than this distributed through intergalactic space in the form probably of neutral hydrogen gas. Bondi (1952, p. 45) estimates that a density as high as $10^{-25}\,\mathrm{g\,cm^{-3}}$ would not conflict directly with any known observations. Bondi (1952) and Hoyle (1949 b) have stated the arguments for a value of ρ of the order $10^{-27}\,\mathrm{g\,cm^{-3}}$.

(*b*) *Age*: *non-steady universes.* With the exception of the steady-state model, all expanding models require the universe to have been in a singular state at a time of the order of T years ago. In general this is a state of infinite density according to the literal interpretation of the theories. It is natural to call the time from that state the *age of the universe* as determined by the model concerned.

It is true that the Eddington–Lemaître model would apparently represent the universe as starting from an Einstein static model, of much higher density than the present value, an infinite length of time ago. But the time since any small finite departure from that special state, which would have to be taken as the effective age, is again of the order of T.

If the expansion phenomenon is of fundamental importance, *the parameter T associated with it must be a characteristic time for the large-scale behaviour of the universe.* If there is anything we can call the 'age' of the universe it could well be expected to be of this order.

About the stage when this was first being appreciated astronomers were independently in process of discovering ages of 10^9 to 10^{10} years for a variety of astronomical systems and of deciding that no known systems demanded ages greater than this (Bok 1946). For instance the age of the oldest rocks in the Earth's crust is estimated by radioactive evidence to be about 3×10^9 years and the age of the oldest stars is estimated from the rate of hydrogen-consumption to be about 5×10^9 years.

All this seemed to fit in well with the notion that the universe as a whole has an age of this order and to provide a strong argument for accepting the implications of the expansion phenomenon. Indeed, it was tempting to suppose that certain of the processes would have been speeded up in their early stages when the universe was in a more congested state. So it was hoped that all the conclusions would ultimately be brought into numerical accord.

This view has since been rejected. No non-steady model with a generally plausible character gives an age of the whole universe much more than 2×10^9

years. On the other hand, the processes that are estimated to have taken two or three times as long require, as is now realized, that the large-scale condition of the universe should not have altered greatly during this longer interval. It is not considered that uncertainties in the observations can account for the discrepancy.

This difficulty has been taken so seriously that the significance of the remarkable agreement as regards orders of magnitude has tended to be lost sight of. Various far-reaching suggestions for resolving the "paradox into which contemporary physics is led", as Milne (1948, p. 224) called it, have been made. Milne's own solution was to use the two time-scales previously described (§5.4). But where he has to use the τ-scale, in which the age of the universe is infinite, there is no explanation of the 'observed' ages of the order 5×10^9 years: where he has to use the t-scale, in which the age of the universe is about 2×10^9 years, the difficulty remains. The same objections apply to a somewhat similar suggestion made by Dirac (1938).

It is to be noted that there is no objection *in principle* to the use of more than one time-scale in physics, save in the case of the steady-state universe. There the PCP implies that there is a time-scale in which the state of the universe is steady and that this is the natural time-scale for *all* physical laws.

(*c*) *Age*: *steady-state universe.* According to the steady-state theory, the average age of the contents of any region at any epoch is $\frac{1}{3}T$. The argument leading to this conclusion can be applied in particular to the ages of the galaxies. But a condensation that is reckoned as a galaxy for this purpose must attain some maturity before it can be recognized observationally as a galaxy. Hence the average age of observed galaxies must be somewhat greater. Any particular galaxy may be of any age, though galaxies several times older than the average must be rare.

Thus we reach the satisfactory position that T is still a characteristic time of the universe as a whole but that it does not impose an upper bound to the age of everything in the universe. From the independent evidence mentioned it is estimated that the age of our own Galaxy is about 5×10^9 years. This is considered to be somewhat older than the average for observed galaxies, and can be accepted as being consistent with the fact that it exceeds the average also in size and probably in mass. In short, the steady-state theory leads us to expect the Galaxy to have an age of the order of magnitude of T but makes no paradox of the fact that it is found to exceed T.

(*d*) *Observable universe.* In the world-maps we have been using the radius of the observable universe at epoch t_0 is cT in all cases where T exists. The mass inside this distance is of the order c^3T/γ in all the cases where it can be related to T. This is about 10^{55} g and is also of the order of the total mass of the universe so frequently discussed by Eddington (1936, 1946).

The steady-state universe is the only model with radius independent of the epoch.

What is more interesting is that the steady-state model and Milne's model are the only ones for which the mass of the observable universe is constant. According to any model other than Milne's, any individual galaxy is radially accelerated or decelerated relative to an observer O. So at some epochs it may have speed less than c and at others speeds greater than c. It can therefore enter or leave the observable universe. This does not happen in Milne's model because

there every galaxy has uniform velocity relative to O. (All this is not in conflict with special relativity because the only space–time of the form (4.17) which can be transformed into that of special relativity is in fact Milne's.) In the steady-state model material is continually disappearing from the observable universe, but in this case it is continually being replaced by newly created matter; so the total mass is conserved.

This non-conservation of total mass in relativistic models in general, as Milne was the first to point out, leaves them open to serious criticism. For it means that a theory that does not profess to treat of creation nevertheless predicts what is equivalent to the creation of matter (or the reverse) in the universe as seen by any observer.

(*e*) *Test of perfect cosmological principle.* The most remote galaxies at present observable are of the order of 10^9 light years away. Therefore we see the corresponding region of the universe as it was 10^9 years ago. According to theories that assign an age of about 2×10^9 years for the universe, there ought to be a considerable difference between the neighbourhood of our own Galaxy seen at its full age and the remote regions seen at only about half that age. There ought to be some observable gradation with distance in the intrinsic features of galaxies or of clusters of galaxies.

According to the steady-state theory, a region seen 10^9 years ago, or at any other epoch, must be statistically the same as a region seen now. So in this case there ought not to be any gradation with distance in the intrinsic features of the galaxies.

This difference ought to furnish an important observational means of deciding between the theories. So far as one knows, with one possible exception no gradation has yet been discovered. In particular, it appears that all known types of galaxy occur at all distances (though the type of galaxy is not recognizable beyond a certain fraction of the maximum distance at which it can be merely detected by a given telescope). The possible exception is the reddening of some galaxies with distance (Stebbins and Whitford 1948). If this effect is real and depends upon *intrinsic* properties of the galaxies, it is a gradation of the sort contemplated, but a surprisingly large one considering the distances of the galaxies concerned. Also it is of the opposite sign to what one would expect it to be if present at all (Bondi 1952, p. 44). If the effect is a *space*-reddening, then it is not immediately relevant. But it could be claimed as favourable to the steady-state theory for a different reason, this being the only theory that positively demands the presence of intergalactic matter.

(*f*) *World-pictures.* We still have not quoted any results obtained from the comparison of theoretical world-pictures with the actual world-picture. We need recall only that Hubble's (1936, 1937) attempts to find a relativistic model giving the best fit gave what he regarded as such generally implausible results that he was ultimately led to doubt the interpretation of the spectral shifts as velocity-shifts. No subsequent work along the same lines as Hubble's has appreciably improved the situation (Heckmann 1942). On the other hand, in work already mentioned, Bondi and Gold got a satisfactory fit with their model, which for this purpose has no adjustable parameters. This was an impressive result, but it is probably inadvisable to emphasize it until more recent data become available.

In due course the work now being done with the new equipment at Mt. Palomar can be expected to provide much more extensive material for such comparisons.

Also, work being done on the characteristics of the nearer galaxies, particularly the 'local cluster', is expected to give much more reliable means of interpreting this material than was available for the older investigations (Baade 1951).

10.3. *Conclusions*

The principle conclusions may be summarized thus: All the different types of theory when applied to the actual universe agree that it cannot be in a static state, and they can all admit the simple interpretation of the observations that it is expanding at the rate expressed by Hubble's law. They all agree in regarding a time of the order of T, the reciprocal of Hubble's constant, as an age of the contents of the universe, being either the time since a unique event of creation or an average time since the operation of a continuing process of creation. They all regard T as determining the size of the observable universe and, generally speaking, they regard it as determining also the mean density and total amount of material in the observable universe.

The following comments upon this situation are called for:

(*a*) Though most of this has been familiar for over twenty years it ought to be remembered that previously almost nothing had been known about the properties of the universe on the scale here concerned. If accepted, it is the greatest advance yet made in our knowledge of the universe. What seems worth emphasizing is that all attempts during these years to approach the subject from fresh viewpoints have tended to reinforce these conclusions as expressed in these very general terms. The controversies occasioned have not affected the situation at this level.

(*b*) The other side of the picture is that this apparently fundamental advance has had astonishingly little effect upon physics and astrophysics in general. We shall notice some possible explanations of this fact.

(*c*) Attention should be given to the leading part played by the parameter T. The significance we place upon this depends upon the status we assign to the CP.

It will be seen that these comments suggest the remaining topics for this review, but their treatment has to be exceedingly brief.

§11. CREATION OF MATTER

11.1. *Unique Creation*

We consider those relativistic models that formally require the universe to have been in a singular state of infinite density at some finite time in the past. Tracing the history of the universe backwards towards this state, the implication is that we reach an earliest state that can be discussed in the language of physics. So far as physics is concerned, the universe must be said to have been *created* in that state. The implication is further that the density of the material was then the greatest known to physics. This means that the whole universe consisted of a mass of *nuclear fluid*. Or, as we may say in Lemaître's (1950) phrase, provided we do not make it signify too much, the universe was created as a 'primeval atom'.

The singularity in the density is associated with a singularity in the rate of expansion specified by the function $V(t)$. So, according to this type of theory, the universe was created in a state of extremely rapid expansion. As soon as the 'primeval atom' came into being it must therefore have proceeded to decay in such a way that, whatever the details of the process, it rapidly became an assembly of ordinary atomic nuclei (and electrons). The continuing rapid

expansion would hinder any subsequent reactions between these nuclei, so that whatever relative abundances they had to begin with would be 'frozen in'. In the case of all elements except hydrogen, helium and the radioactive series these should be the relative abundances observed in the actual universe.

The possibility of *a theory of the abundances of the elements* based on these ideas has been investigated very fully by nuclear physicists (see review by ter Haar, 1950). In spite of some promising particular results, it has not proved possible to account for the observed abundances throughout the whole range of nuclear species. Since the investigations appear to have been exhaustive as regards the processes that could have occurred, this failure constitutes a serious criticism of this type of cosmology.

11.2. *Continual Creation*

The steady-state theory, as we have seen, requires the continual creation of new matter at the rate $3\rho/T$ per unit volume per unit time, i.e. about 3×10^{-44} g cm^{-3} sec^{-1}. Since the calculations are made for a model of the smoothed-out universe, this is to be interpreted in the first instance merely as a 'smoothed-out' rate. If the theory is applied to the actual universe, it does not show whether we are to suppose the newly created matter to appear uniformly throughout space or only, say, in the vicinity of galaxies. But *if the circumstances of the appearance of the fresh matter are conditioned in any way by the state of the matter already present it is almost impossible to account for this state being steady.*

This is perhaps the strongest reason for the form of the hypothesis of continua creation put forward by the authors of the steady-state theory, though they and Hoyle are able to advance a number of other reasons. This hypothesis may be expressed as: *Atomic hydrogen is continually appearing throughout space as a result of a process that is random and spontaneous.*

The randomness ensures statistical uniformity; it and the spontaneity ensure independence of the state of matter already present. It is not essential that the initial product should be neutral hydrogen; it might be a neutron, or a proton and electron, but it is an essential feature of the hypothesis that the effective result should be predominantly a continually renewed supply of hydrogen gas. Its postulated rate of appearance according to the value quoted above is about one atom per cm^3 per 10^{12} years, which is utterly beyond any means of direct detection.

According to this theory, as at present formulated, the relative abundances of the chemical elements have therefore no immediate relation to the creation process.

A particularly attractive feature of the theory is, however, that it makes the whole state of the physical universe depend explicitly upon an atomic process. Using Hoyle's model, there is only one fundamental parameter T for the large-scale behaviour of the universe and this is directly determined by the rate of creation.

§ 12. GALAXIES

We have seen that cosmological theories have inevitably to concern themselves with the creation of matter. We have also seen that effectively all the observable matter in the universe is that contained in the galaxies. Obviously, therefore, a very important problem is that of the way in which the created matter becomes formed into the concentrations which we recognize as these galaxies. The

solution of this problem must provide the first link between cosmology and ordinary astrophysics. Since little progress has yet been made with it, no surprise need be occasioned by the fact that cosmological investigations have hitherto had such small influence upon progress in astrophysics.

All cosmologies require the notion that the matter now forming galaxies should at some earlier stage have been distributed in space in a more uniform and tenuous state. They consider that the galaxies are the result of some form of gravitational instability in this state.

Non-steady-state theories. Until recently these theories have had very little to say on the subject except to suggest qualitatively that there could be a stage in the expansion of the universe at which the density and speed of dispersion would have reached values favourable to the operation of gravitational instability (Lemaître 1948, 1950, 1951). Von Weizsäcker (1951) has lately outlined a much more elaborate theory in which turbulence in the initial medium plays an essential part along with expansion and gravitational instability. It is too soon to say whether quantitative success is forthcoming.

Steady-state theory. The essential difference between the problem here and in other theories is that in the steady-state theory there never is a stage in which the 'medium' is devoid of galaxies. The problem is therefore to account for the perpetuation of the population of galaxies; the theory contemplates no origination of this population. This was briefly explained by McCrea (1951 b). The tendency of the expansion process is to remove galaxies from any given region of space, and the tendency of the creation process is to occupy this region with a uniform distribution of hydrogen gas. The presence of the galaxies must however prevent the actual establishment of a uniform distribution. The gas near a galaxy already present must fall into it under gravitational attraction, and that galaxy must therefore grow by accretion. The gas not predominantly under the gravitational influence of any single galaxy will tend to form condensations under its own gravitational instability aided by the perturbing effect of the combined action of surrounding galaxies. Such condensations will then grow by accretion and form new galaxies. Apart from a few 'order of magnitude' estimates (Bondi and Gold 1948, Hoyle 1948, 1949 a, McCrea 1950) this theory too has not been worked out. If successful, it should evaluate the mean distance between galaxies, the mass of a galaxy of given age, and the ratio of galactic to intergalactic matter in terms of the rate of creation of matter.

Special attention should be called to the inference that a galaxy is continually accreting hydrogen from outside itself. In Hoyle's (1949 b, 1950) view some such process is demanded to account for the predominance of hydrogen in our Galaxy. If so, this would be additional observational evidence in favour of the theory.

Cosmological principle. The topics of this section concern the importance of the ***non-uniformity*** of the universe over times and distances of a certain size. We may associate this size with what, for the moment, we may call the 'small-scale' behaviour of the universe. It has now to be admitted that we have no absolutely secure estimates of the relative magnitudes of quantities associated with the behaviour of the actual universe (*a*) on the small-scale, (*b*) in the region accessible to present observations, (*c*) on the large-scale. This profoundly affects the status of the CP in the work.

Let us use the terms *a*-, *b*-, *c*-quantities. If *a*-quantities are small compared with *b*-quantities, then it is reasonable to speak of testing by observation whether

the actual universe obeys the CP or the PCP and, if so, to proceed tentatively with the application of theories based on these principles. If *a*-quantities are not small compared even with *c*-quantities, then non-uniformity is an essential feature of the actual universe and we cannot treat it using theories that assign to the CP the status it has in the theories we have described. Nevertheless, if *a*-quantities are not small compared with *b*-quantities, an intermediate position is possible. It might then be possible, as a working approximation, to treat the *b*-region as uniform and, moreover, to compare its behaviour with the predictions of a theory that, merely for simplicity, uses a model uniform throughout. In that case the CP would be a simplifying assumption for the theory and an approximation for its application, and need have no fundamental significance for cosmology.

It will be clear from what has been said in § 10 that there is at present no demand for degrading the CP. But the safest prediction we can make about the results of future observations is that they will reveal unexpected phenomena. Probably the next safest is that the general features of the theories discussed here will be found nevertheless to retain some range of validity. For the history of physical science shows that a serious major development in physical thought has probably never had to be rejected outright.

§ 13. APPENDIX ON CURRENT DEVELOPMENTS

At the meeting of the International Astronomical Union in Rome in September 1952 a preliminary account was given of work that may lead to a considerable revision of the quantitative observational results quoted in the foregoing article, which had already been written in its present form. Until the astronomer chiefly concerned has published his findings, it is scarcely proper to quote him explicitly. But it may be stated that the other astronomers present appeared to be prepared to accept the implications of his work.

A crucial step in finding the distance-scale of the galaxies is the determination of the zero-point of the period–luminosity relation for the variable stars known as ' classical cepheids '. This fixes the scale-factor for all the other distances. It now appears that the zero-point hitherto employed may have been considerably in error. The new determination would have the consequence of at least *doubling all previous estimates* of the distances of external galaxies (but would not call for a revision of distances *within* our own Galaxy). In particular, this would imply that the value of the parameter T should be about 4×10^9 years rather than about 2×10^9 years, provided the other measurements involved in determining the value of T are assumed to have been correct.

The observational results briefly quoted in the text of the present report, apart from one or two features explicitly stated to be more recent, have stood for the past fifteen years or more. It is certainly disconcerting to learn that such a drastic revision may be required at almost a single stroke. Nevertheless, the following comments seem to be permissible:

(*a*) While the possible revision is known to arise from observations of the highest quality, astronomers in general have not yet had the opportunity to examine the evidence in detail. Therefore the result cannot yet be described as being 'accepted'.

(*b*) Much other independent new work has been in progress for the re-determination, in particular, of Hubble's constant. This work requires the re-determination of other quantities besides those involved in the re-determination of the distance-scale. Its results are now awaited with great interest.

(*c*) As mentioned in the text, it has recently been appreciated that there are probably many more intrinsically faint galaxies than had previously been estimated. This means that certain statistical estimates of distance may have been over-estimated. If so, this would tend to counterbalance the proposed upward revision of the distance-scale. Thus it is possible that the revision of the overall picture of the observed universe may be less drastic than seems at first sight to be implied.

(*d*) The present article is concerned mainly with the aims and methods of cosmological theory and not with the detailed comparison of theoretical and observational results. In particular, if the conclusions tend to be in favour of some form of steady-state theory, this is because of the general considerations advanced rather than because it has yet been shown that such a theory gives the best detailed agreement with observation. If the empirical value of the parameter T is actually about twice what had previously been estimated, then it is true that the difficulties of the non-steady-state theories are reduced. But this does not affect the positive arguments that seem to support the steady-state theory.

It must be repeated that the observational results quoted in the paper are those that were in use *before* the proposed possible revision which is the occasion for this appendix.

REFERENCES

ALPHER, R. A., and HERMAN, R. C., 1950, *Rev. Mod. Phys.*, **22**, 153.

BAADE, W., 1951, *Publications of Observatory of Univ. of Michigan*, **10**, 7.

BOK, B. J., 1946, *Mon. Not. R. Astr. Soc.*, **106**, 51.

BONDI, H., 1948, *Mon. Not. R. Astr. Soc.*, **108**, 104 ; 1952, *Cosomology* (Cambridge : University Press).

BONDI, H., and GOLD, T., 1948, *Mon. Not. R. Astr. Soc.*, **108**, 252.

DIRAC, P. A. M., 1938, *Proc. Roy. Soc.* A, **165**, 199.

EDDINGTON, A. S., 1930, *Mon. Not. R. Astr. Soc.*, **90**, 668; 1936, *Relativity Theory of Protons and Electrons* (Cambridge: University Press); 1946, *Fundamental Theory* (Cambridge: University Press).

EINSTEIN, A., 1951, *The Meaning of Relativity*, 5th edn. (London : Methuen).

TER HAAR, D., 1950, *Rev. Mod. Phys.*, **22**, 119.

HECKMANN, O., 1942, *Theorien der Kosmologie* (Berlin : Springer).

HOYLE, F., 1948, *Mon. Not. R. Astr. Soc.*, **108**, 373 ; 1949 a, *Ibid.*, **109**, 365 ; 1949 b, *Nature, Lond.*, **163**, 196; 1950, *The Nature of the Universe* (Oxford : University Press).

HUBBLE, E., 1936, *Realm of the Nebulae* (Oxford : University Press); 1937, *The Observational Approach of Cosmology* (Oxford: University Press); 1951, *Proc. Amer. Phil. Soc.*, **95**, 461.

JORDAN, P., 1947, *Die Herkunft der Sterne* (Stuttgart: Hirze); 1952, *Schwerkraft und Weltall* (Brunswick : Vieweg).

KERMACK, W. O., and MCCREA, W. H., 1933, *Mon. Not. R. Astr. Soc.*, **93**, 519.

LEMAÎTRE, G., 1948, *Acad. roy. Belgique*, Cl. Sci., **34**, 551, 1950, *The Primeval Atom* (New York : Van Nostrand), 1951, *Acad. roy. Belgique*, Cl. Sci., **37**, 291.

MCCREA, W. H., 1935, *Z. Astrophys.*, **9**, 290 ; 1950, *Endeavour*, **9**, 3 ; 1951 a, *Proc. Roy. Soc.* A., **206**, 562 ; 1951 b, *Trans. Victoria Institute* **83**, 105.

MCCREA, W. H., and MILNE, E. A., 1934, *Quart. J. Math.* (Oxford ser.), **5**, 73.

MACH, E., 1893, *The Science of Mechanics* (London: The Opera Court Publishing Company).

MCVITTIE, G. C., 1939, *Proc. Phys. Soc.*, **51**, 529.

MILNE, E. A., 1935, *Relativity, Gravitation and World-structure* (Oxford: University Press); 1948, *Kinematic Relativity* (Oxford: University Press); 1952, *Modern Cosmology and the Christian Idea of God* (Oxford : University Press).

ROBERTSON, H. P., 1933, *Rev. Mod. Phys.*, **5**, 62 ; 1935, *Astrophys. J.*, **82**, 284 ; 1936, *Ibid.*, **83**, 187, 257.

SCIAMA, D. W., 1953, *Mon. Not. R. Astr. Soc.*, **113**, in the press.
SHANE, C. D., and WIRTANEN, C. A., 1950, *Proc. Amer. Phil. Soc.*, **94**, 13.
DE SITTER, W., 1917, *Mon. Not. R. Astr. Soc.*, **78**, 3; 1932, *Kosmos* (Harvard: University Press).
STEBBINS, J., and WHITFORD, A. E., 1948, *Astrophys. J.*, **108**, 413.
TOLMAN, R. C., 1934, *Relativity, Thermodynamics and Cosmology* (Oxford: University Press).
WALKER, A. G., 1936, *Proc. Lond. Math. Soc.*, (2) **42**, 90; 1944 a, *Observatory*, **65**, 242; 1944 b, c, *J. Lond. Math. Soc.*, **19**, 219, 227.
VON WEIZSÄCKER, C. F., 1951, *Astrophys. J.*, **114**, 165.
WEYL, H., 1923, *Phys. Z.*, **24**, 230.
WHITTAKER, E. T., 1942, *Proc. Roy. Soc., Edinburgh*, **61**, 160.

The following are some books on the subject not referred to in the text:

COUDERC, P., 1952, (tr. Sidgwick, J. B.) *The Expansion of the Universe* (London: Faber and Faber).
EDDINGTON, A. S., 1933, *The Expanding Universe* (Cambridge: University Press).
JOHNSON, M., 1948, *Time, Knowledge and the Nebulae* (London: Faber and Faber).
MCVITTIE, G. C., 1949, *Cosmological Theory*, 2nd edn. (London: Methuen).
WHITROW, G. J., 1949, *The Structure of the Universe* (London: Hutchinson).
WHITTAKER, E. T., 1942, *The Beginning and End of the World* (Oxford: University Press)

THE NEW UNSTABLE COSMIC-RAY PARTICLES

By G. D. ROCHESTER and C. C. BUTLER
The Physical Laboratories, University of Manchester

CONTENTS

PAGE

I. Introduction ... 365
II. Neutral unstable particles ... 366
§ 1. Introduction ... 366
§ 2. Typical experimental arrangements ... 366
2.1. General features ... 366
2.2. Counter control systems ... 367
§ 3. The production of V^0-particles ... 367
§ 4. Decay products ... 368
4.1. Heavily ionizing secondary particles ... 368
4.2. Penetration of the secondary particles through matter ... 370
4.3. Differential momentum spectra ... 370
4.4. Number of decay products ... 372
4.5. Summary ... 373
§ 5. Q-values and masses of V^0-particles ... 373
5.1. Dynamics of two-body decay processes ... 373
5.2. Manchester data ... 375
5.3. Pasadena data ... 378
5.4. Other Q-value data ... 380
5.5. The masses and decay schemes of V^0-particles ... 381
§ 6. The mean lifetimes of V^0-particles ... 382
§ 7. The neutral zeta meson ... 382
III. Charged unstable particles ... 383
§ 1. Introduction ... 383
§ 2. The τ-meson ... 383
2.1. Introduction ... 383
2.2. Primary particles ... 383
2.3. Secondary particles ... 384
2.4. Mass of the τ-meson ... 385
§ 3. K-mesons ... 385
3.1. Introduction ... 385
3.2. κ-mesons and χ-mesons ... 386
§ 4. S-particles ... 387
§ 5. Charged V-particles ... 388
5.1. Introduction ... 388
5.2. Slow charged V-particles ... 388
5.3. Fast charged V-particles ... 389
5.4. The double decay ... 391
§ 6. Charged zeta particles ... 391
§ 7. Decay schemes for K-mesons and charged V-particles ... 392
§ 8. The interaction of K-mesons with matter ... 395
§ 9. Production of charged heavy mesons ... 395
9.1. Introduction ... 395
9.2. The creation of heavy mesons in 'jets' ... 396
9.3. Energy balance in 'jets' ... 399
9.4. The production of τ- and K-mesons ... 399
9.5. The production of charged V-particles ... 399
§ 10. Lifetimes of heavy mesons ... 400
10.1. Introduction ... 400
10.2. The τ-meson ... 400
10.3. Charged V-particles ... 401
IV. Summary of the decay schemes of neutral and charged unstable particles ... 402
V. Future outlook ... 404

ABSTRACT. A detailed description is given of the properties of the new unstable particles with masses between 400 m_e and 2200 m_e. Wilson cloud chamber experiments have provided definite evidence for the existence of a heavy neutral particle, the V_1^0-particle, of mass 2190 m_e, which decays into a proton and a negative π-meson. A somewhat lighter neutral particle, known as a V_2^0-particle, also exists, and possible decay schemes are discussed. The properties of the charged V-particles observed in Wilson cloud chambers are compared with the three main types of charged unstable particles found in photographic emulsions. The three types are the τ-meson, of mass 977 m_e, which decays into three charged π-mesons, the κ-particle, of mass about 1100 m_e, which decays into a μ-meson, and two or more neutral particles, and the χ-meson, of mass (900–1500) m_e, which decays into a charged π-meson and one neutral particle. The available data indicate that some of the S-particles found in cloud chambers are identical with the χ-particles, and that the charged V-particles probably consist of a mixture of κ- and χ-particles. The very slight evidence for the existence of charged particles with masses about 500 m_e is also described. A discussion is given of many of the possible decay schemes of the numerous particles, their frequencies of production and lifetimes.

I. INTRODUCTION

It is the purpose of this report to review what is known of the existence and properties of the new unstable particles, that is, particles which decay spontaneously and have masses greater than the π-meson. The fact that a review of this length can be devoted entirely to this subject is striking evidence of the immense advances which have been made in the past three years. In Powell's excellent review of the properties of μ- and π-mesons in these *Reports* in 1950, the possible existence of the new forms of unstable particle was noted, but the evidence was based on relatively few cases. It is not surprising that at that time many physicists had doubts about their reality and significance. The position has now completely changed. The main types of neutral and charged heavy meson have been completely confirmed and new types have been discovered by the further application of the two most powerful techniques in this field, the cloud chamber and the electron-sensitive photographic emulsion. The cloud chamber has continued to provide almost the whole of the evidence for the neutral particles, while the photographic emulsion has given the most precise information about the charged particles.

It is now clear that the whole subject is one of great complexity owing to the many different types of particle which lie in relatively narrow ranges of masses and lifetimes and yet have entirely different modes of decay. An extensive new field of particle physics has been opened up which will engage the attention of physicists for many years to come. The conclusions reached in this report may be somewhat tentative, but it is hoped that the experimental facts are correct. No attempt will be made to discuss the theoretical implications of the various decay schemes.

The two main classes of particle, the neutral and the charged, will be treated in sections II and III. The neutral particles will be termed V^0-particles (after Blackett and Anderson), and clear evidence will be presented for the existence of a particle, termed V_1^0, which has a mass of about 2190 m_e and decays into a proton and a π-meson. The evidence for the existence of other, lighter, neutral particles, termed V_2^0 and ζ^0, and of masses about (900–1000) m_e and (500–600) m_e respectively, will also be examined.

The evidence relating to the many charged particles will be given in section III. It is now known that in the mass range (900–1500) m_e there occur the S-particles,

the τ- and κ-mesons, and possibly the χ-meson, each with a characteristic mass and form of decay. Where κ- and χ-mesons occur together, and cannot be distinguished, they will be termed K-mesons, and where all the types (and any others, at present unknown, which may lie in the same range of masses) have to be referred to as a class they will be termed B particles or barons [B ≡ 'baron', Greek 'heavy', after Menon (private communication)]. The term charged V-particle will also be used since it describes a well defined empirical type. It will be shown that this type is probably similar to the K-particle. Also considered in this section will be the charged particles, termed ζ-particles, lying in the mass range (400–700) m_e.

An attempt will be made in sections IV and V to summarize the present state of our knowledge and to indicate possible modes of attack for the future.

The early history of the subject has been given by Powell (1950), Rochester and Rosser (1951), Butler (1951), Rochester and Wilson (1952), and Rossi (1952).

II. NEUTRAL UNSTABLE PARTICLES

§1. INTRODUCTION

Neutral V-particles were discovered in the course of a study of penetrating showers with a cloud chamber by the authors (Rochester and Butler 1947). A copy of the original photograph, with its characteristic V-shaped track or fork, is shown in Plate I. Detailed consideration of this photograph led to the conclusion that the fork was produced by the spontaneous decay in flight of a neutral particle, of minimum mass about 800 m_e, into two singly charged particles. This interpretation was confirmed by Seriff *et al.* in 1950 and subsequently by many other workers in various parts of the world. In this section of the report we summarize the present position concerning these particles. The possible existence of a new type of neutral particle, the ζ^0-particle, is also discussed (Danysz *et al.* 1952).

§2. TYPICAL EXPERIMENTAL ARRANGEMENTS

2.1. *General Features*

With the exception of one possible observation of a V^0-fork in the photographic emulsion by Hopper and Biswas (1950), all the known V^0-particles have been found by the counter-controlled cloud chamber. The main data relating to the cloud chambers used in the study of this problem are given in table 1. It will be noted that many hundreds of V^0-particles have now been found, and that most groups of workers operate at mountain altitudes where the rate of V^0-particles is at least ten times that at sea level. Magnetic fields are used in four of the experiments. A field is important not only for the determination of the sign of the particles—if the direction is known—but also for the determination of the momentum.

While the use of a magnetic field is a very desirable feature of cloud chamber work, important results have been obtained without fields. This is especially true of the large multiplate chambers (see table 1). The use of many thin plates permits the determination of momentum by scattering and of energy by range, and affords a very effective means of studying the interaction with matter of both primary and secondary particles.

2.2. *Counter Control Systems*

Since V^0-particles were discovered in penetrating showers, all groups of workers have used counter systems designed to select these showers as efficiently as possible. A typical counter system, used by the Pic-du-Midi group, is shown schematically in fig. 1. In this arrangement two combinations of counters are used, the first requiring two particles at both trays A and B, and the second, one particle at tray A and two at each of B and C. The second combination is used to increase the probability of detecting nuclear interactions due to primary ionizing particles in the layers of lead immediately above the chamber, and also

Table 1. Cloud Chamber Data

Group	Location and altitude (m)	Chamber dimensions (cm)	Metal plates in chamber	Magnetic field (gauss)	Approx. no. of V^0-tracks
(1)	(2)	(3)	(4)	(5)	(6)
Armenteros *et al.* (1951 a, b)	Pic-du-Midi, France 2867	Diam. 28 Illtd. depth 7	2 cm Pb or 1·5 cm W–Cu alloy	7500	200
Astbury *et al.* (private communication)	Jungfrau, Switzerland 3460	54×54. Illtd. depth 8	2·5 cm Pb	6000	80
Leighton *et al.* (1951, 1953)	Pasadena 220 Mt. Wilson California 1750	Two chambers each 33×19 high. Illtd. depth 12, used one above the other	2·5 cm Pb between chambers	5000	134
Thompson *et al.* (1951)	Bloomington, Indiana Sea level	Diam. 28 Illtd. depth 7·5	—	3100	—
Bridge and Annis (1951) Bridge *et al.* (1952, 1953 a, b)	Echo Lake, Colorado 3000	50×50 Illtd. depth 15	11 plates of Pb each 0·6 cm thick	—	75
Deutschmann (1952)	Göttingen Sea level	75×75 30 deep	4 Pb 1·8 cm thick, 1 paraffin wax plate 9 cm thick, and 2 graphite plates 5 cm thick	—	7 (5 from Pb, 1 from paraffin)
Fretter *et al.* (1953)	Berkeley, California Sea level Lake Sabrina, California 2750	51 wide×61 high ×23 deep	7 Pb each 1·2 cm	—	75

in the plate inside the chamber. It is estimated that at least 80% of the counts are due to nuclear interactions of average energy (20–40) KMEV [i.e. (2–4) × 10^{10} EV] occurring in the lead. The counting rate and the number of V^0 tracks obtained in a typical 'run' at the Pic-du-Midi are given in table 2.

§ 3. THE PRODUCTION OF V^0-PARTICLES

Up to the present no detailed examination of the showers with which V^0-particles are associated has been made, but it is known that they are produced by both neutral and charged primary particles in almost all types of penetrating shower (Fretter *et al.* 1953). Plate V is an example of a V^0-particle associated with a large shower, and Plate II one associated with a small shower. In showers of average energy, about 30 KMEV, the frequency of production of

V^0-particles is probably a few per cent of the frequency of π-mesons. Fretter *et al.* (1953) found one V^0-decay per 18 identified penetrating showers and one per 80 penetrating particles (see also table 2).

Leighton *et al.* (1953) have recently given the results of a study of the origins of 134 V^0-particles, and they find that, of 37 V^0-particles produced in the 2·5 cm lead plate between their two chambers, 27 V^0-particles are associated with interactions produced by charged particles which have themselves come from interactions in the lead *above* the upper chamber (see table 1). Of the remaining ten V^0-particles, six are produced by charged primary particles and four by

Table 2. Rate of V^0 Production

Arrangement	$\begin{cases} A_2B_2 \\ A_1B_2C_2 \end{cases}$
No. of photographs	3923
Time (h)*	490
No. of V^0-particles	30
No. of penetrating particles per V^0	83 ± 20

* Corrected for resetting time.

neutral particles. Since the majority of shower particles at this energy seem to be π-mesons, it follows that energetic π-mesons can produce V^0-particles. This point still needs to be proved, however, for heavy mesons are also produced in these interactions (see III, §9).

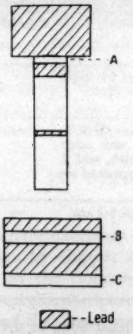


Fig. 1. Arrangement used by the Pic-du-Midi Group.

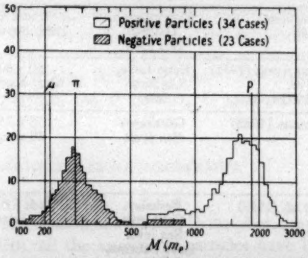


Fig. 2. The mass spectrum of the secondary particles. (*After* Leighton *et al.* 1953). (All measured values represented in the diagram are plotted as rectangles of the same area, whose widths correspond to the mass ranges derived from the central values of the measured momenta and the extreme limits of the ionization estimates.)

§4. DECAY PRODUCTS

V^0-particles cannot be studied directly, but much valuable information can be obtained by an examination of the tracks produced by the charged decay products.

4.1. *Heavily Ionizing Secondary Particles*

Seriff *et al.* (1950) found several secondary particles with specific ionizations considerably above the minimum value, but they were not able to make any mass measurements. Hopper and Biswas (1950) and Hopper (private communication) found a V-shaped track in a photographic emulsion which may have been produced

PLATE I. The first photograph of the decay of a neutral V-particle. The V^0-decay is the wide-angle fork a few millimetres below the plate to the right-hand side of the main shower. The arms of the fork are indicated by arrows. The magnetic field was 3500 gauss.
(*After* G. D. Rochester and C. C. Butler)

PLATE II. A V^0-particle produced in a very simple interaction. A positively-charged particle, $p \sim 2$ k MeV/c, stops in the 1 cm lead plate and a V^0-decay appears about 0·5 cm below the plate. The plane of the fork intersects the line of flight of the primary particle at the centre of the plate. (*After* R. Armenteros *et al.*)

PLATE III. An example of a V^0-decay in which one of the secondary particles can be identified as a proton. The fork is to the left-hand side of the chamber and the proton is represented by the track at the left of the fork. The other track, which is at minimum ionization, was probably formed by a π-meson.
(*After* R. Armenteros *et al.*, 1951 a)

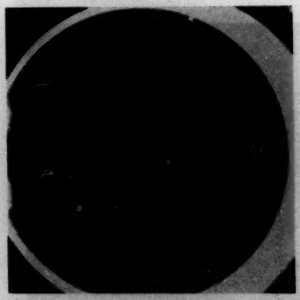


PLATE IV. An example of a V^0-decay in which the positive particle (represented by the track to the left of the fork) has a momentum of (270 ± 30) MeV/c and yet ionizes close to minimum, an observation which proves it cannot be a proton. The ionization of a proton would be 10 times minimum. It is, indeed, unlikely that the mass of this particle was greater than the π-meson. The negative particle could not be identified. The magnetic field was 3100 gauss. (*After* R. W. Thompson *et al.*)

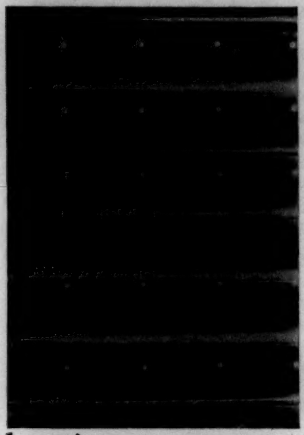


PLATE V. A striking photograph which shows that neutral V-particles originate in high-energy nuclear interactions. The nuclear explosion is initiated in the fourth lead plate by a high-energy charged particle, and among the products is a V^0-particle which decays below the fifth plate and lies towards the left-hand side of the chamber. The arms of the fork are indicated by arrows. There was no magnetic field. (*After* W. B. Fretter)

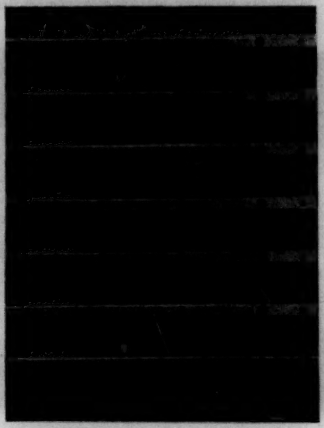


PLATE VI. An unusual photograph of a V^0-decay in which both charged secondaries interact in a 0·25 in. lead plate. This event probably represents the decay of a V_1^0-particle into a proton and a π-meson. (*After* H. S. Bridge *et al.*)

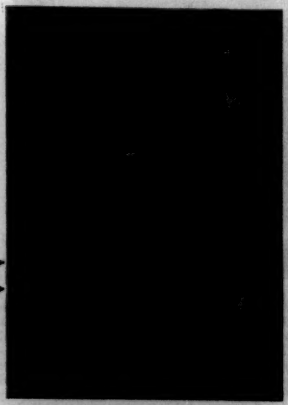


PLATE VII. The unusual case of a V^0-decay which has originated in a low-energy star in the gas. This follows from the fact that, to within experimental error, the plane through the tracks of the secondary particles contains the centre of the star.

(*After* J. P. Astbury *et al.*)

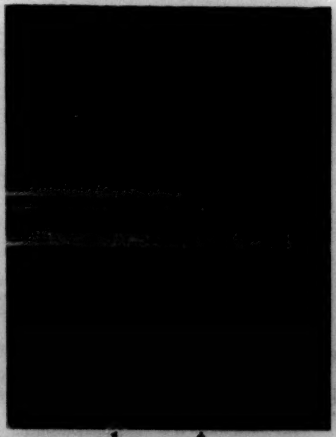


PLATE VIII. A V^0-decay in which the secondary particles can be identified. The decay occurs just below the plate. The positive particle, which has a mass of the order of the proton, is to the right of the fork, and the negative particle, which has a mass of (230–430) m_e, and is probably a π^--meson, is to the left. The magnetic field was 5000 gauss.

(*After* R. B. Leighton *et al.*)

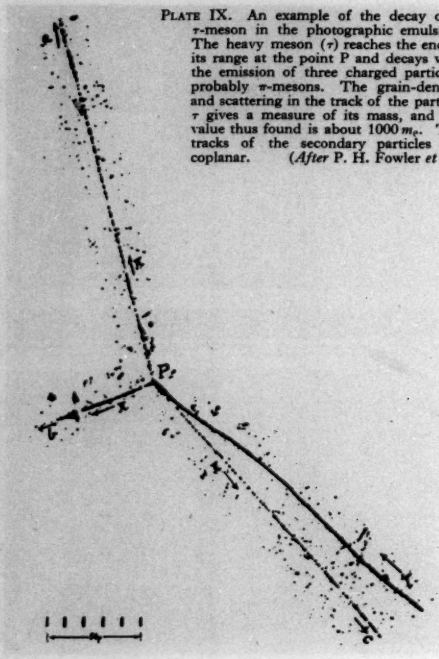


PLATE IX. An example of the decay of a τ-meson in the photographic emulsion. The heavy meson (τ) reaches the end of its range at the point P and decays with the emission of three charged particles, probably π-mesons. The grain-density and scattering in the track of the particle τ gives a measure of its mass, and the value thus found is about $1000\,m_e$. The tracks of the secondary particles are coplanar. (*After* P. H. Fowler *et al.*)

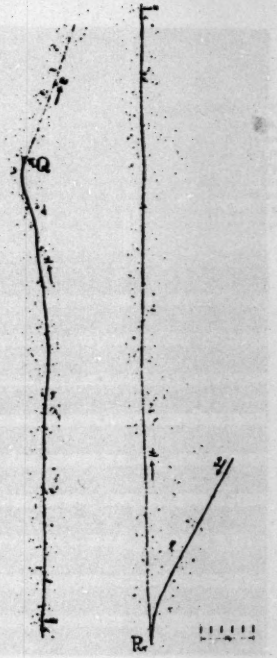


PLATE X. A heavy meson, κ_2, reaches the end of its range at a point P and decays with the emission of a μ-meson of range $1098\,\mu$. The μ-meson stops at the point Q and emits an electron. The track of the κ-meson is given in two parts. The mass of the particle as given by measurements of scattering and grain-density is $(1125 \pm 230)\,m_e$.

(*After* C. O'Ceallaigh)

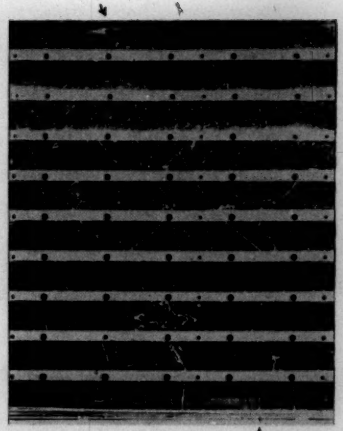


PLATE XII. An example of a typical S-particle. The primary particle (indicated by an arrow) enters at the top of the chamber and after penetrating six 0·25 in. lead plates stops in the seventh plate and apparently decays to a charged secondary particle whose track is at minimum ionization. The secondary particle (also indicated by an arrow) penetrates three lead plates without interaction.
(*After* M. Annis *et al.*)

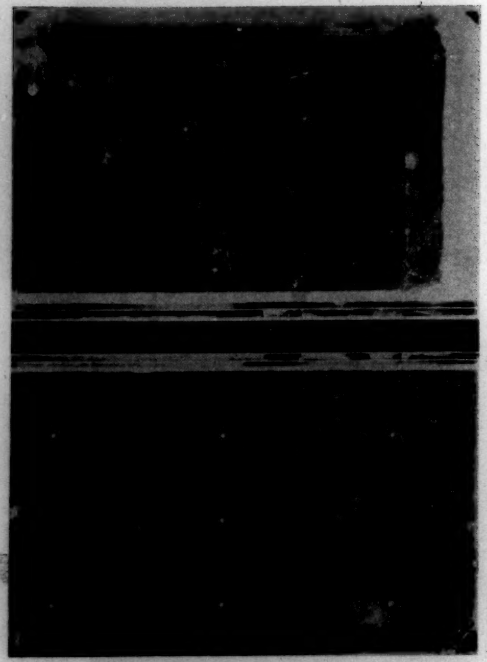


PLATE XI. An example of the decay of a τ-meson in the cloud chamber. The magnetic field was 5000 gauss.
(*After* R. B. Leighton and S. D. Wanlass)

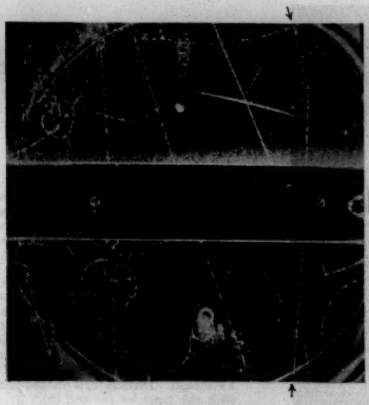


PLATE XIII. A photograph of a penetrating shower containing a charged V-particle. The V-decay is indicated by arrows, and consists of a positive particle which decays at the top right-hand side of the cloud chamber into a positive secondary particle. The charged secondary particle penetrates the 3 cm lead plate without interaction. The magnetic field was 7200 gauss.

(*After* G. D. Rochester and C. C. Butler.)

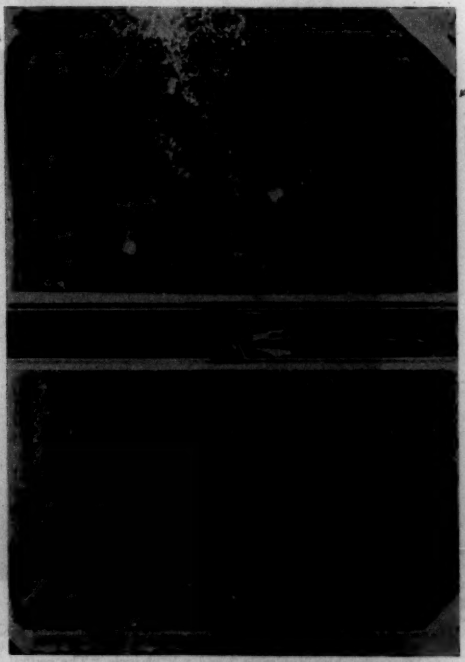


PLATE XIV. A cloud-chamber photograph of the decay of a slow charged V-particle. The charged V-particle comes down from the right-hand side and decays just above the plate to form a secondary particle which moves up to the left. The magnetic field was 5000 gauss.

(*After* R. B. Leighton and S. D. Wanlass)

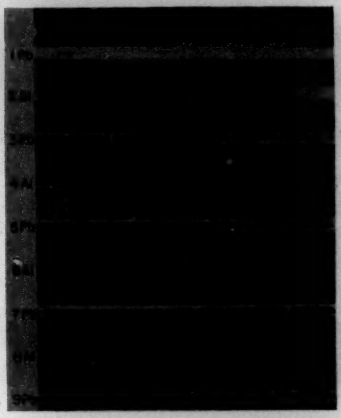


PLATE XV. An example of the decay of a charged V-particle to a secondary particle which seems to be a π-meson. The primary charged V-particle comes from an interaction in the third plate (0·26 in. lead), passes through a 0·31 in. plate of aluminium, and then decays in the gas of the chamber to form particle (a). This particle is scattered in the aluminium plate and emerges as particle (b) which slows down and finally stops in the ninth plate. The measurements of scattering and range of particle (b) lead to the conclusion that it is a π-meson. Particle (c) is unconnected with the particles (a) and (b).

(*After* H. S. Bridge and M. Annis)

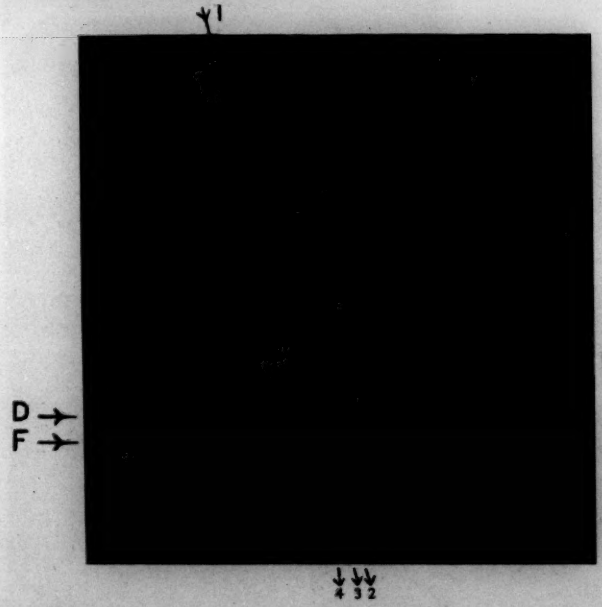


PLATE XVI. The double decay (see fig. 12). This photograph shows an event in which a charged particle entered the chamber at 1 and after penetrating the lead plate decayed at D into the charged particle 2 and a V^0-particle. The neutral V-particle subsequently decayed at F into two charged particles which produced tracks 3 and 4.

(*After* R. Armenteros *et al.*)

PLATE XVII. 'Jets' produced by primary particles of energy greater than 1000 kMev. The nomenclature $x+yp$ indicates an event containing y shower particles and x heavily-ionizing particles.

(*After* R. R. Daniel *et al.*)

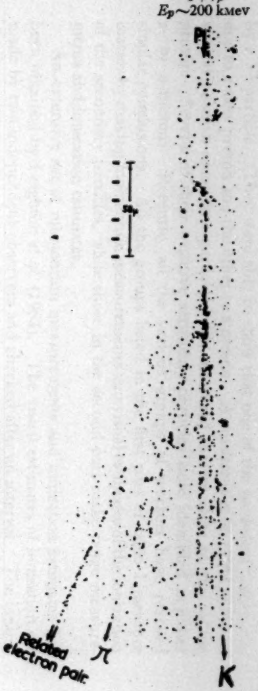


PLATE XVIII. An example of a 'jet' which provides evidence for the creation of K-particles. The tracks of a π-meson and a K-particle are marked. Also indicated is an electron pair which has presumably come from one of the photons produced by the decay of a π^0-meson.

(*After* R. R. Daniel *et al.*)

by a V^0-particle. They were able to show that one of the secondary tracks was due to a slow proton and the other to a meson. Armenteros *et al.* (1951 a) found several positive secondary particles with measured masses close to that of the proton. Several of their negative secondary particles were identified as either π- or μ-mesons. Recently, all the various groups of workers have reached similar conclusions. All the groups listed in the first part of table 1 measure masses by making momentum measurements and visual estimates of the ionization of the secondary particles, while those in the second part make range measurements and ionization estimates.

An extensive set of momentum measurements and ionization estimates has been obtained by Leighton *et al.* (1953). The visual estimates of ionization are made by two independent observers and checked after an interval. The ionization of each secondary particle is compared with that of a standard track on the same photograph. Convenient standard tracks are produced by particles whose ionizing power is at the minimum value. Leighton *et al.* define minimum ionization as the ionization of a penetrating particle whose momentum exceeds 1 kMeV/c. Alternatively they use electrons with momenta in the range 50–100 MeV/c whose ionizations are assumed to be 1·3 minimum, thus allowing for the rise in the ionization of energetic electrons. The experiments of Ghosh, Jones and Wilson (1952) and Carter and Whittemore (1952) show that there is also a rise in the ionization of mesons with momenta above 500 MeV/c. In fact the ionization of many of the standard tracks produced by penetrating particles may have been as high as (1·1–1·3) minimum. Leighton *et al.* claim to be able to estimate ionizations within the range once to twice minimum as well as above twice minimum; they do not quote errors of measurement but give ranges inside which they consider the true ionizations almost certainly lie.

Using the above technique, Leighton *et al.* have made a preliminary classification of the masses of the secondaries produced by 87 V^0-particles. It was impossible to make mass determinations for all the secondaries, so they were put in the categories shown in table 3.

Table 3. Classification of the Masses of Secondary Particles of 87 V^0-Tracks

	(1) $m_-<700\,m_e$	(2) $m_->700\,m_e$	(3) m_- indeterminate
(*a*) $m_+<700\,m_e$	2	1	4
(*b*) $m_+>700\,m_e$	42	0	3
(*c*) m_+ indeterminate	29	0	6

The forty-two positive particles in category (1*b*) are heavier than the corresponding negative secondaries. Furthermore, many of the 29 particles in category (1*c*), which are all substantially at minimum ionization, must be more massive than their corresponding negative particles; this conclusion follows from the relative numbers of particles in categories (3*a*) and (1*c*), which should be equal for any decays with positive and negative secondaries of equal mass. Leighton *et al.* consider that their measurements indicate that at least 80% of the positive secondaries are more massive than the corresponding negative particles. There is definite evidence, however, that a small number of V-tracks have positive secondaries which are less massive than the proton; these events are listed in table 3 in categories (1*a*), (2*a*) and (3*a*).

Leighton *et al.* have made mass measurements on some secondary particles with ionizations above the minimum value; the mass values are shown in fig. 2 in the form of histograms. Most of the positive particles are grouped round the proton mass and the negative particles are distributed symmetrically about the π-meson mass. It therefore seems certain that many of the V^0-particles decay into protons and negative π-mesons; these unstable particles are known as V_1^0-particles. Leighton *et al.* consider two of the heavily ionizing mesons were probably μ-mesons and that a small number of the secondaries had masses of about $1000\,m_e$. In view of the width of the two distributions around the π-meson mass and the proton mass it is impossible to identify with confidence a small number of μ-mesons or positive particles with masses about $1000\,m_e$ among so many π-mesons and protons.

Armenteros *et al.* (private communication) have measured the momenta of both secondaries of 40 V^0-particles. The ionizations of twenty of the positive secondaries were equal to, or above, twice the minimum ionization. The experimental values of the masses of these particles are distributed about the proton mass, and Armenteros *et al.* conclude that the particles were protons. Three of the negative secondaries were mesons.

Typical V^0-tracks with heavily ionizing secondaries are shown in Plates III and VIII.

4.2. *Penetration of the Secondary Particles through Matter*

All groups of workers have reported that the secondary particles readily penetrate considerable thicknesses of lead and occasionally make nuclear interactions. Thus Bridge *et al.* (1952) have photographed a decay in which both secondary particles penetrated several lead plates and then each produced a small nuclear interaction: this unusual photograph is reproduced in Plate VI and shows conclusively that both the secondaries are nuclear interacting particles. Bridge *et al.* (1953 a) have reported a value of $(163 \pm 40)\,\mathrm{g\,cm^{-2}}$ of lead for the interaction length for the secondaries of V^0-particles. Clearly the secondaries could be protons or π-mesons, but the accuracy of this measurement is such that it is impossible to exclude a small fraction of μ-mesons or other non-interacting particles.

4.3. *Differential Momentum Spectra*

The differential momentum spectra of the secondary particles were first measured by Armenteros *et al.* (1951 b); they concluded that most of the negative secondaries have much lower momenta than the positive secondaries, a conclusion which is even more marked in the results of Leighton *et al.* (1953) (see fig. 3). Armenteros *et al.* (1951 b) found, however, a number of lightly ionizing positive particles with momenta less than 700 Mev/c. Upper estimates of the masses of these particles can be made by assuming that the ionizations were as high as twice minimum; the results for three accurately measured corresponding positive and negative particles are given in table 4. The error on each upper estimate is obtained from the uncertainty in the momentum measurement.

The positive particles in table 4 cannot be definitely identified, but it seems very unlikely that they were protons. Thus Armenteros *et al.* (private communication) find that at least 10% of their V^0-forks do not have proton secondaries.

The simplest interpretation of the data of table 4, in terms of common particles, is that both positive and negative particles are π- or μ-mesons, but it should be stressed that so far there is no *direct* evidence in support of this assumption. An example of a V^0-track with a positive secondary lighter than the proton is reproduced in Plate IV.

Table 4. Measurements on Three Associated Positive and Negative Secondary Tracks

(1) Event number	66	171	198
(2) Momentum of positive secondary (Mev/c)	300 ± 40	485 ± 60	300 ± 24
(3) Upper estimate of mass of positive particle (assuming $I/I_{min}=2$) (m_e)	750 ± 50	1200 ± 200	700 ± 50
(4) Momentum of negative secondary (Mev/c)	120 ± 30	610 ± 140	3400 ± 1500
(5) Upper estimate of mass of negative particle (assuming $I/I_{min}=2$) (m_e)	325 ± 50	1500 ± 300	>1800

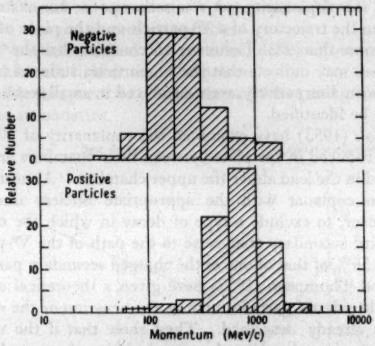


Fig. 3. The differential momentum spectra of the secondary particles. (*After* Leighton *et al.* 1953)

Leighton *et al.* have also obtained V^0-tracks with positive secondaries less massive than the proton; they were classified in categories (1*a*), (2*a*) and (3*a*) in table 3. It seems likely that several of these events are similar to those described in table 4. One of them is of particular interest; the momentum of the positive particle is 190 ± 20 Mev/c and the ionization less than 1·5 times minimum; thus the mass is less than about $350\, m_e$. The momentum of the corresponding negative particle is 210 ± 40 Mev/c and its ionization from two to four times minimum, which indicates that the mass of the particle is in the range $(450\text{–}1000)\, m_e$. Leighton *et al.* examine the possibility that the event was produced by the decay of a positive particle and conclude that there is no evidence in favour of this explanation. Therefore they suggest that the photograph shows that some V^0-particles decay into light positive secondaries—probably π-mesons—and negative secondaries heavier than the π-meson. A study of the data in table 4 shows that at least one of the events (66) probably differs from this type, since the mass of the negative secondary is close to the mass of the π-meson.

Provisionally, all the V^0-particles which produce positive secondaries lighter than the proton are known as V_2^0-particles.

4.4. *Number of Decay Products*

It is known that each V^0-particle decays into two charged particles, and there is no direct evidence for the presence of neutral secondary particles. Several groups of workers have examined their photographs carefully for the cascade showers which might result from γ-rays produced directly or from π^0-mesons occurring among the decay products, but none has reported finding any evidence for their existence. If there are only two secondaries the plane of each V-track must also contain the point where the nuclear interaction in which the V^0-particle was produced took place. Fretter *et al.* (1953) have tested the coplanarity of 50 V^0-tracks, and of these 43 forks were found to be coplanar within the accuracy of measurement; the remaining seven forks were apparently non-coplanar. The accuracy of this coplanarity test was rather poor; for example, the error in the angle between the trajectory of a V^0-particle and the plane of the two secondaries was often more than 2°. Fretter *et al.* consider that the seven apparently non-coplanar cases may indicate that the V^0-particles suffered nuclear scattering in the lead plates or, alternatively, were produced in small secondary interactions which could not be identified.

Leighton *et al.* (1953) have examined the coplanarity of 60 V^0-particles, of which 37 were produced in the plate between their chambers and the remaining 23 were produced in the lead above the upper chamber. Almost all of the decays could have been coplanar with the appropriate nuclear interactions. It is impossible, however, to exclude a type of decay in which the momentum component of the third secondary transverse to the path of the V^0-particle is on the average less than 50% of that of one of the charged secondary particles.

Brueckner and Thompson (1952) have given a theoretical discussion of the influence of a light neutral particle, such as the neutrino or the π^0-meson, on the coplanarity tests already described. They show that if the velocity β of the V_1^0-particle is approximately unity, then the deviations from coplanarity are small and probably cannot be detected. They calculate the distribution of the quantity $\delta' = \delta\beta(1-\beta^2)^{-1/2}$, where δ is the angle between the path of the V^0-particle and the plane containing the charged secondaries, and β is the velocity of the V^0-particle. The velocity is calculated approximately using the vector sum of the momenta of the two charged secondaries and an average mass for the V_1^0-particle. It is necessary to know the Q-value of the decay process, and the following were chosen to fit in with the experimental Q-values (see II, §5.5):

$$\left.\begin{aligned} &V_1^0 \rightarrow p^+ + \pi^- + \text{neutrino} + 130\ \text{MeV} \\ \text{and} \quad &V_1^0 \rightarrow p^+ + \pi^- + \pi^0 + 70\ \text{MeV}. \end{aligned}\right\} \qquad \ldots\ldots(\text{II.1})$$

Leighton *et al.* have compared their experimental values of δ' with the theoretical distribution for schemes (II.1). Most of their V_1^0-particles have β approximately 0·7. Either of the three-body schemes could explain the coplanarity data although there is a tendency for the experimental observations of δ' to be smaller than the values predicted by the theory. In order to reach a definite conclusion it is necessary to restrict attention to low-energy V_1^0-particles for which the departure from coplanarity is more marked if schemes (II.1) are correct.

Recently Bridge *et al.* (1953 a) have tested the coplanarity of 15 V_1^0-tracks. Low energy events were chosen with values of β approximately 0·2. Very careful coplanarity tests have been made and a high accuracy has been achieved since the angle between the axes of the two cameras was 60°, a very much larger angle than that used by either Fretter *et al.* or Leighton *et al.* The experimental histogram of the values of δ' is significantly sharper than the form predicted by Brueckner and Thompson. Thus Bridge *et al.* conclude that their data can best be interpreted by a two-body scheme rather than by the three-body schemes (II.1). We cannot conclude, however, that these results prove that all the V_1^0-particles decay into only two charged secondary particles; it is possible that a small fraction has more complicated decay schemes.

4.5. *Summary*

(*a*) Many of the positive secondary particles are protons and the corresponding negative secondary particles are probably π-mesons. The V^0-particle responsible for this type of V^0-track is termed the V_1^0-particle. Most V_1^0-particles decay into only the two charged secondary particles.

(*b*) Some V^0-particles, called V_2^0-particles, decay into two charged secondaries, both of which are lighter than the proton.

(*c*) It is possible that particles of mass about 1000 m_e and μ-mesons occasionally occur among the secondaries.

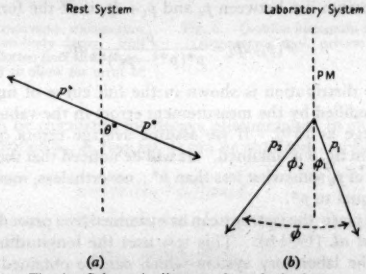


Fig. 4. Schematic diagram of two-body decay.

§ 5. *Q*-VALUES AND MASSES OF THE V^0-PARTICLES

Following the conclusions reached in § 4, the data have been analysed on the assumption that the only decay products are the two charged secondary particles. It is impossible to present a single analysis because the different groups of workers have not published all the basic data from which calculations can be made. Before discussing the details of the various results already obtained it is useful to give a brief summary of the dynamics of two-body decay processes.

5.1. *Dynamics of Two-body Decay Processes*

The behaviour of the two secondary particles in the rest system of the V^0-particle is particularly simple since they must be emitted with equal and opposite momenta p^* as shown in fig. 4. It is assumed that they are emitted isotropically in the rest system. The momentum p^* depends only on the mass M of the

V^0-particle and the masses m_1 and m_2 of the secondaries, and it can be obtained from the relativistic energy equation which may, for the rest system, be written in the form $M=(p^{*2}+m_1^2)^{1/2}+(p^{*2}+m_2^2)^{1/2}$, where the masses are expressed in Rossi units.

In the laboratory frame of reference the two secondary particles have momenta p_1 and p_2 and their tracks are inclined to one another at an angle ϕ. The relativistic energy equation for this frame of reference has the form

$$(M^2+P^2)^{1/2}=(m_1^2+p_1^2)^{1/2}+(m_2^2+p_2^2)^{1/2} \qquad \ldots\ldots(\text{II}.2)$$

where P, the momentum of the V^0-particle, is given by

$$P^2=p_1^2+p_2^2+2p_1p_2\cos\phi. \qquad \ldots\ldots(\text{II}.3)$$

Thus M can be calculated from measurements of p_1, p_2 and ϕ provided that the masses of the secondary particles are known.

An examination of the degree of isotropy of the secondary particles in the rest system of the V^0-particles can be made in the following way. In the first place the component of p^* transverse to the path of the V^0-particle, namely p_t, is unaffected by the transformation from the rest system to the laboratory system. For a particular V^0-track the value of p_t can be calculated from the relation

$$p_t=p_1p_2\sin\phi/P. \qquad \ldots.(\text{II}.4)$$

Armenteros *et al.* (1951 *b*) showed that the differential probability that the transverse momentum lies between p_t and p_t+dp_t is of the form

$$\mathscr{P}(p_t)\,dp_t=\frac{p_t}{p^*(p^{*2}-p_t^2)^{1/2}}\,dp_t. \qquad \ldots\ldots(\text{II}.5)$$

The form of the distribution is shown in the full curve of fig. 5. In practice, however, it is modified by the measurement errors in the values of p_t, which are sometimes as large as 15%. If we assume average errors of this magnitude, the dotted curve in fig. 5 is obtained. It will be noticed that the peak of the curve occurs at a value of p_t somewhat less than p^*; nevertheless, most of the values are approximately equal to p^*.

In the second place, the isotropy can be examined by a procedure also suggested by Armenteros *et al.* (1951 b). This test uses the longitudinal components of momentum in the laboratory system which can be obtained from the corresponding momentum components in the rest system by means of Lorentz transformations. A quantity α is defined such that

$$\alpha=(p_{1L}-p_{2L})/(p_{1L}+p_{2L}), \qquad \ldots\ldots(\text{II}.6)$$

where p_{1L} and p_{2L} are the longitudinal components of p_1 and p_2. For a two-body decay scheme it can be shown that:

$$\alpha=\frac{p_1^2-p_2^2}{P^2}=\frac{m_1^2-m_2^2}{M^2}+2p^*\cos\theta^*\left(\frac{1}{M^2}+\frac{1}{P^2}\right)^{1/2}, \qquad \ldots\ldots(\text{II}.7)$$

where θ^* is the angle between the path of the V^0-particle and the line of flight of the two secondary particles in the rest system (see fig. 4). Since the secondary particles are emitted isotropically in the rest system, $\cos\theta^*$ takes all values equally from $+1$ to -1. Thus the average of a set of α-values (α^*) is given by

$$\alpha^*=(m_1^2-m_2^2)/M^2,$$

which is a constant for a particular decay scheme. Furthermore, it will be seen that if α is plotted against $1/P$ all the values lie inside two limiting hyperbolae given by (II.7) when $\theta^*=0$ and π. For a given value of P the experimental values of α should be evenly distributed between the limiting hyperbolae. In addition to the limiting hyperbolae, others can be drawn corresponding to various values of θ^*. Now $p_t=p^*\sin\theta^*$, so that these hyperbolae are lines of constant transverse momentum. All these properties of the $(\alpha, 1/P)$ relation are shown in fig. 8.

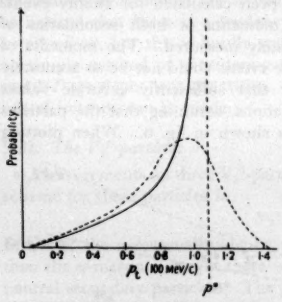


Fig. 5. Theoretical transverse momentum distribution for two-body decay. Full line, eqn. (II.5); dotted line, same equation, but modified to allow for error of measurement.

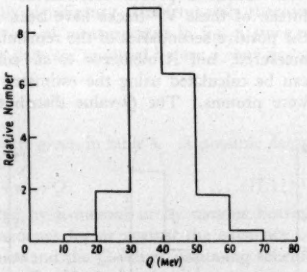


Fig. 6. Q-value histogram for 20 V_1^0-particles. (Armenteros *et al.*, private communication.)

Values of α can be calculated for decays photographed in chambers without magnetic fields provided ϕ_1, ϕ_2 (see fig. 4) can be obtained, since it can readily be shown that

$$\alpha=[\sin(\phi_2-\phi_1)]/\sin\phi. \qquad \ldots\ldots(\text{II.8})$$

These values of α can conveniently be plotted against the non-dimensional quantity $2p_t/P$ where

$$\frac{2p_t}{P}=\frac{2\sin\phi_1\sin\phi_2}{\sin\phi}. \qquad \ldots\ldots(\text{II.9})$$

Lines of constant θ^* can be plotted on the same diagram and isotropy of emission in the centre-of-mass system can be tested.

5.2. *Manchester Data*

Armenteros *et al.* (1951 b and private communication) have made useful measurements on about 40 V^0-tracks out of a total of 200. In each case the momenta p_1 and p_2 and the opening angle ϕ were measured. Twenty of the V^0-tracks have heavily ionizing positive secondary particles which were probably protons and therefore can be interpreted by the decay of V_1^0-particles (see II, § 4.1). Three of the positive secondaries were lighter than the proton; they were produced by the decay of V_2^0-particles (see II, § 4.3). The secondaries of the remaining seventeen events cannot be identified and therefore these V^0-particles cannot be classified directly.

5.21. *The V_1^0-particles.*

It was shown in II, §4.1 that the charged secondary particles produced by the decay of V_1^0-particles are almost certainly protons and negative π-mesons. Following the work of Bridge *et al.*, described in II, §4.4, it is reasonable to assume that no neutral secondary particles are produced. Thus the most likely decay scheme is

$$V_1^0 \rightarrow p + \pi^- + Q, \qquad \ldots\ldots(II.10)$$

where Q is the kinetic energy released in the rest system of the V_1^0-particle, so that $Q = M - (m_1 + m_2)$. Values of Q have been calculated for twenty events with identified proton secondaries. The momenta of both secondaries of fifteen of these V_1^0-tracks have been accurately measured. The momenta of the positive secondaries of the remaining five events could not be so accurately measured, but Armenteros *et al.* consider that sufficiently accurate values can be calculated using the estimated ionizations, assuming that the particles were protons. The Q-value distribution is shown in fig. 6. When plotting

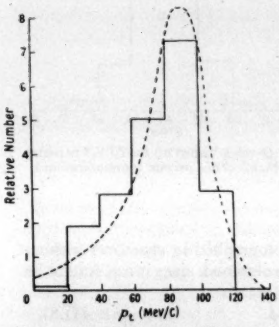


Fig. 7. p_t distribution for 20 V_1^0-particles. The dotted curve is the theoretical distribution for a two-body decay scheme with $p^* = 98$ MeV/c.

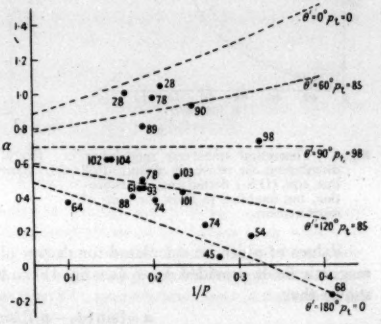


Fig. 8. $(\alpha, 1/P)$ diagram. The values of the transverse momentum in MeV/c are given next to each point on the diagram.

Note. θ' is the θ^* of the text.

this histogram the events were 'spread' in the cells according to the estimated errors in the Q-values. For example, if an event occurs at 38 ± 10 MeV, it counts as 0·1 of an event in the cell 20–30 MeV, 0·5 of an event in the cell 30–40 MeV and 0·4 of an event in the cell 40–50 MeV, i.e. the true Q is assumed to lie evenly in the range $(38-10)$ MeV to $(38+10)$ MeV, and is spread among the three cells accordingly. The best Q-value is 40 ± 3 MeV and the width of the distribution (half-width about 10 MeV) probably arises solely from the individual measurement errors, which range from 8% to 32%. The distribution is probably slightly asymmetrical, having a 'tail' towards the high Q-values. This is to be expected since, for a given percentage accuracy in the momentum measurements, the percentage error in the Q-value increases approximately proportionally to Q.

The degree of the isotropy of emission of the two secondary particles in the centre-of-mass system can be examined in two ways. The observed histogram of the transverse momentum components p_t is shown in fig. 7. Now the average

error on the values of p_t is only about 10%; assuming this value, the theoretical distribution can be fitted to the experimental histogram (see II, §5.1 and fig. 5). The theoretical distribution is shown as a dotted curve in fig. 7. The value obtained for p^* is 98 MeV/c, corresponding to a Q-value of 37 MeV for scheme (II.10) which is consistent with the value already found. The second test of isotropy is made using the $(\alpha, 1/P)$ relation given in eqn. (II.7). The data for the twenty V_1^0-particles are plotted in fig. 8, where the lines of constant p_t have been drawn for a Q-value of 37 MeV, assuming scheme (II.10). The values of the transverse momenta are shown beside each point on the diagram. It will be noticed that there is a tendency for the observations to lie below the line $\alpha = \alpha^* = 0{\cdot}69$, but this is hardly statistically significant. Unfortunately there is no simple method of considering the errors of measurement in fig. 8 since α and $1/P$ are dependent quantities. More data are required before a convincing test of isotropy can be made.

5.22. *The V_2^0-particles.*

Measurements on three V_2^0-particles were given in table 4. A possible decay scheme for these particles is

$$V_2^0 \rightarrow \pi^+ + \pi^- + Q. \qquad \ldots\ldots (II.11)$$

Either of the π-mesons might be replaced by μ-mesons or by mesons heavier than the π-meson. As yet there is no evidence for or against the existence of neutral secondary particles. The p_t values and the Q-values, assuming scheme (II.11), for the three V_2^0-particles are given in table 5.

Table 5. Values of p_t and Q for Three V_2^0-Particles (assuming $V_2^0 \rightarrow \pi^+ + \pi^- + Q$)

Event no.	66	171	198
p_t (MeV/c)	108	229	35
Q (MeV)	114	260	245
Standard error in Q (MeV)	16	55	100

The amount of data is very meagre and, furthermore, the accuracy of the Q-values is not very high. Two of the Q- and p_t-values are consistent with a Q-value of about 250 MeV. It is possible that there are several V_2^0-particles with different decay schemes and Q-values.

5.23. *The unclassified V^0-particles.*

Seventeen V^0-particles have positive secondary particles with relativistic velocities and momenta close to or greater than 700 MeV/c. It is, therefore, impossible to classify them directly into V_1^0- and V_2^0-particles. The data can be examined, however, by assuming first that all these decays were due to V_1^0-particles. The seventeen Q-values obtained fall into two groups, those less than 100 MeV and those greater than 100 MeV. The first group contains eleven cases; histograms of the Q- and p_t-values are very similar to those shown in figs. 6 and 7. We conclude that these eleven V^0-particles were probably V_1^0-particles with a Q-value close to 40 MeV. The Q-values for the second group, consisting of six events, are completely inconsistent with a unique Q-value of about 40 MeV, assuming scheme (II.10). Two possible explanations are now considered, firstly that these events were produced by V_1^0-particles with Q-values

greater than 100 MeV, and secondly that they were not produced by the decay of V_1^0-particles at all.

The six values of p_t may be compared with the values of p^* calculated using the individual values of Q and the masses m_1 and m_2 of the decay scheme (II.10). They are given in columns (3) and (4) of table 6. It can easily be shown that the probability of a value of the transverse momentum being greater than a

Table 6. Data on Six Possible V_1^0-Particles

(1)	(2)	(3)	(4)	(5)	(6)
53	290	185	341	84	184 ± 50
90	194	143	292	88	120 ± 16
125	225	207	290	71	230 ± 65
130	285	195	339	82	200 ± 78
149	113	170	190	46	208 ± 50
203	363	140	399	94	129 ± 30

(1) Event no.; (2) Q (MeV) assuming scheme (II.10); (3) observed p_t (MeV/c); (4) calculated p^* (MeV/c); (5) $\mathscr{P}'$ (%); (6) Q (MeV) assuming scheme (II.11).

chosen p_t is given by $\mathscr{P}' = 100(1 - p_t^2/p^{*2})^{1/2}$. The calculated values of $\mathscr{P}'$ are given in column (5) of table 6. Now $\mathscr{P}'$ can be greater than 80% for a single observation but hardly greater than 80% for four out of six cases. Furthermore, if the events are due to the decay of V_1^0-particles with various Q-values greater than 100 MeV then a study of the α-values shows that in nearly all these cases the proton was emitted upwards in the centre-of-mass system. It is possible that this type of event is preferentially selected for measurement because an upward moving proton will have a momentum in the laboratory system which can probably be measured. However, if the six events described in table 6 were produced by V_1^0-particles, then the values of $\mathscr{P}'$ show that among the unmeasured photographs there should be about 40 V_1^0-tracks of comparable quality (i.e. track length, freedom from obvious distortion, etc.) with high momentum positive secondaries and much lower momentum negative secondaries. These events would have values of p_t of about 300 MeV/c. Armenteros *et al.* have not found this large number of events among their unmeasured forks. In view of these results, they consider that the six forks described in table 6 were probably not produced by the decay of V_1^0-particles.

The six events under discussion could also be interpreted by the V_2^0-particle decay scheme (II.11); the corresponding Q-values are given in column (6) of table 6. Four of the values are remarkably close to the last two given in table 5 and, considering the consistency of the data, it is very probable that these V^0-forks were produced by the decay of V_2^0-particles with a Q-value within the range 175–250 MeV. The values of the transverse momenta are also consistent with this Q-value. Events Nos. 66, 90 and 203 of tables 5 and 6 give much lower Q-values than the remaining six events. They can also be interpreted by the decay of another type of V_2^0-particle into a τ-meson and a π-meson (see II, §5.3).

5.3. *Pasadena Data*

The measurements made by Leighton *et al.* on the masses of the secondary particles show that almost all their V^0-tracks have positive secondaries with masses greater than $700\,m_e$ and thus can be interpreted by scheme (II.10).

Confirmatory evidence was obtained by plotting a histogram of 92 α-values, 62 of these values being obtained from measurements of the momenta p_1 and p_2 and the remaining 30 from measurements of either p_1 or p_2 and the angles ϕ_1 and ϕ_2. Very few of the α-values were negative, and even some of these cases consisted of positive secondaries heavier than the π-meson. A few cases, however, had unidentified positive secondaries and also negative values of α. This number of the cases is doubled to obtain an estimate of the number of possible V_2^0-decays. The result shows that not more than 15% of the 92 cases could be due to the decay of V_2^0-particles.

Leighton *et al.* have completely measured the secondary tracks of 38 V^0-particles with positive secondaries heavier than $700\,m_e$ and negative particles lighter than $700\,m_e$. These decays constitute almost all of those in category (1 *b*) of table 3, and are interpreted by means of scheme (II.10). The Q-values range from 10 MeV to 89 MeV, and their distribution is shown in the cross-hatched portion of the histogram of fig. 9. Q-values for an additional group of 25 cases

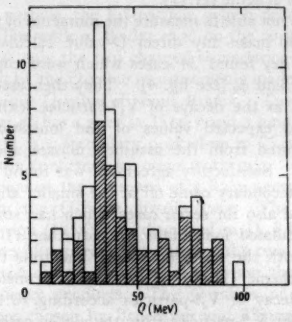


Fig. 9. The Q-value distribution for 62 cases obtained by Leighton *et al.* (1953).

are also plotted in fig. 9. This class consists of V^0-particles whose positive secondaries could not be identified, and are included in category (1 *c*) in table 3. All these V^0-particles have values of α greater than $+0{\cdot}5$ and are unlikely to be produced by the decay of V_2^0-particles. The histogram shown in fig. 9 was obtained by counting the number of events with measured Q-values in each 5 MeV range without regard to the individual errors of measurement. Although they vary considerably around an average of about 30%, several of the best cases have errors of only 15%. The histogram appears to have two maxima, a very marked one at 35 ± 3 MeV and a smaller one at 75 ± 5 MeV; about 25% of the selected V_1^0-decays seem to belong to the higher Q-value group. In view of the large errors in the individual Q-values, the second peak is not statistically significant. These individual errors indicate that the total data could be consistent with the average Q-value of 46 ± 5 MeV, but this average is considerably displaced from the main peak. It seems more plausible to take the view either that the distribution has a peak at 35 MeV and a 'tail' stretching up to about 100 MeV or possibly that there are two peaks.

Finally, Leighton *et al.* consider a third group of seven decays which cannot be interpreted by the decay of V_1^0-particles. Three of these could, when interpreted by scheme (II.11), give Q-values of about 110 MeV. One of the remaining decays appears to have a positive π-meson secondary and negative secondary of mass (450–1000) m_e. If the negative secondary is assumed to be a τ-meson (see III, § 2), then the Q-value of this type of neutral V-particle is approximately 60 MeV.

5.4. *Other Q-Value Data*

Apart from the main results described above, data of smaller statistical weight have been obtained by Thompson *et al.* (1951) and Fretter *et al.* (1953). Thompson and his co-workers found two cases which, when interpreted according to the scheme (II.10), gave Q-values of 28 ± 10 MeV and 31 ± 5 MeV respectively. A third case could not be fitted into this scheme since the positive particle had a momentum of 270 ± 30 MeV/c and yet was lightly ionizing. A proton of this momentum would have had an ionization of ten times minimum. This event can be interpreted by scheme (II.11).

Fretter *et al.* were not able to measure the momenta of the secondary particles and hence could not make any direct Q-value calculations using equations (II.2) and (II.3). They found 34 cases which were coplanar and were thus able to determine ϕ_1 and ϕ_2 (see fig. 4). They then assumed that these forks could be interpreted as the decays of V_1^0-particles with masses in the range (2200–2400) m_e. The expected values of the ionizations of the secondary particles were computed from the assumed masses and compared with the observed ionizations. Satisfactory agreement was found for 16 cases in which the heavily ionizing secondary came off at the smaller angle from the direction of the V^0-particle and also for seven cases which had secondaries at minimum ionization. The calculated ionizations for a further 11 cases, however, were totally inconsistent with the observed values, and thus these events could not be interpreted by scheme (II.10). Fretter *et al.* consider that they can be interpreted by the decay of V_2^0-particles according to scheme (II.11). The estimated masses of the V_2^0-particles from the eleven cases is about 800 m_e.

The same workers have attempted to determine the mean Q-value for their best 18 low-energy V_1^0-particles. It has already been shown that the ionizations of the secondaries are consistent with Q-values for the V_1^0-particles in the range (50–150) MeV. The best Q-value was found by examining the distributions of the angles of emission θ^* of the secondaries in the rest system for several Q-values within the acceptable range. The best Q-value is the one which gives a distribution of θ^* close to the expected form for isotropic emission of the secondaries in the rest-system. The value obtained by this procedure is 125 MeV, which is much higher than the values of about 40 MeV obtained by the Manchester and Pasadena workers. It can easily be shown that if the method used by Fretter *et al.* is to be reliable, then events with values of the opening angle ϕ between 0° and 180° must be included in the analysis. There must be no experimental bias against measuring events with ϕ greater than 90°. In practice there is, however, a marked tendency to make the best measurements on forks with ϕ less than 90°, and it can be shown that this form of selection bias tends to give a high Q-value when the method of Fretter *et al.* is used.

Recently Bridge *et al.* (1953 a) have made a number of Q-value measurements on low-energy V_1^0-particles using data obtained in their multiplate chamber.

The forks are accurately coplanar with their origins inside the chamber (see II, §4.4). From the behaviour of the secondary particles in the chamber it is possible to decide which is the proton and which is the meson. In addition, the range of one of the secondaries, usually the meson, can be measured in twenty cases. Using these data, and assuming scheme (II.10), Q-values for the twenty events can be obtained. With two exceptions all these Q-values are consistent with a unique Q-value in the range (35–40) MeV. Rossi *et al.* consider that the measurements on the two exceptional cases are not very reliable, and they conclude that their limited amount of data is consistent with a unique Q-value for the V_1^0-particle. They have also obtained five forks which they consider can be interpreted satisfactorily by means of scheme (II.11).

5.5. *The Masses and Decay Schemes of V^0-Particles*

Despite the considerable amount of data on V^0-particles, the present position regarding the masses and decay schemes is somewhat confused. It is established that there is a neutral particle, the V_1^0-particle, which decays into a proton and a negative π-meson. The work of Bridge *et al.* on the coplanarity of low-energy V_1^0-particles shows that only the two charged secondaries are produced. This conclusion is confirmed by the Q-value measurements made by Armenteros *et al.* Both groups find a unique Q-value of about 40 MeV. The Q-value distribution obtained by Leighton *et al.* has a peak at 35 MeV and a long 'tail' up to 100 MeV. It is possible that there is a secondary peak close to 75 MeV, but the existence of this peak has not been confirmed by other workers.

Further evidence against the presence of a light neutral secondary particle can be obtained from the Q-value distributions. The theoretical work of Brueckner and Thompson (1952) was discussed in II, §4.4. They have also calculated a distribution of apparent Q-values for the two schemes (II.1). The apparent Q-value is that obtained on the assumption that the charged secondaries are the only decay products. The Q-values of 70 MeV and 130 MeV for schemes (II.1) were chosen by Brueckner and Thompson in order to give their theoretical distributions maxima at about 35 MeV, which is close to the maxima of the experimental apparent Q-value distributions. Leighton *et al.* have compared their experimental results with the calculated ones, and the fit is very poor. Similarly the fit is very poor indeed for the results of Armenteros *et al.*

We may conclude that there is strong evidence that the majority of the V_1^0-particles decay into two charged secondary particles, a proton and a negative π-meson, and that the kinetic energy released is between 35 MeV and 40 MeV. Thus the mass of the unstable particle is $(2190 \pm 5)\,m_e$. It is possible that a small fraction of the V_1^0-decays are more complex.

All groups find evidence for a lighter neutral particle, the V_2^0-particle, which seems to decay into light mesons. Up to the present the decay schemes of these particles have not been settled because only upper estimates of the masses of the secondaries have been obtained and, therefore, no identifications can be made. Armenteros *et al.* consider that some of their V_2^0-particles probably decay into two charged π-mesons. The mass of this type of V_2^0-particle is about $950\,m_e$. It is probable that there is a second type of V_2^0-particle which decays into a π-meson and a heavy meson, possibly a τ-meson. Armenteros *et al.* and Fretter *et al.* find relatively more V_2^0-particles than Leighton *et al.* Very little

significance can be attached to this result because the various groups select their data for analysis and make their measurements in different ways and also use different Geiger-counter arrangements and cloud chambers.

§6. THE MEAN LIFETIMES OF V⁰-PARTICLES

Fretter *et al.* have measured the mean lifetime of their V_1^0-particles using the method described in III, §9, and they find a minimum value of $3{\cdot}6 \times 10^{-10}$ sec; the most probable value appears to be about 10^{-9} second. In contrast Leighton and Alford (1952) report provisionally that the lifetime of V_1^0-particles is $2{\cdot}5 \times 10^{-10}$ second. The difference between the two results may possibly be explained by the fact that Fretter *et al.* measured their times of flight from the origin of the V⁰-particles, whereas they should have been measured from fiducial surfaces inside the illuminated region of the chamber (see III, §9). Recently, Astbury (private communication) has obtained a value of about $1{\cdot}6 \times 10^{-10}$ sec for the lifetime of the V_2^0-particle.

Both Fretter *et al.* and Leighton *et al.* have considered whether V⁰-particles are produced in pairs. A small percentage of the photographs showing the decay of V⁰-particles contains a second decay associated with the same penetrating shower as the first decay. Fretter *et al.* find only three cases of pairs out of a total of 65 pictures containing V⁰-particles. Both groups of workers conclude that there is no evidence for the production of V⁰-particles in pairs.

§7. THE NEUTRAL ZETA MESON

Recently Danysz *et al.* (1952, 1953), using the photographic emulsion technique, have made some observations which indicate that a small fraction of the π-mesons ejected from nuclear disintegrations (energies 2–20 Mev) are associated both in direction and in energy. This evidence is based on (*a*) a study of the relative energies of pairs of measurable π-mesons ejected at small angles to one another, and (*b*) a statistical analysis of the angular distribution of the shower particles from stars of type 2p.

Cosyns (1949, 1950) and Mei and Pickup (1952) have made an analysis of the angular distribution of shower particles from stars of higher multiplicity, $n_s \sim 10$, and have also found evidence for the occurrence of a greater number of pairs of particles, making a small angle with one another, than are expected on statistical grounds. Similar studies by Daniel *et al.* (1952), Kaplon and Ritson (1952), and Lal *et al.* (1952 a, b), at very high energies, yield little or no evidence for the occurrence of such pairs.

It was tentatively suggested by Danysz *et al.* that the simplest interpretation of the data was in terms of the existence of a very short-lived meson ζ^0 which decayed according to the scheme,

$$\zeta^0 \rightarrow \pi^+ + \pi^- + Q, \qquad \ldots\ldots (II.12)$$

where Q is of the order of a few Mev. Brueckner and Watson (1952) have suggested that an effect such as that observed could be accounted for in terms of an interaction between pairs of π-mesons. Danysz *et al.* (1953) have come to the conclusion that with the present experimental data it is not possible to distinguish between these two hypotheses. Experiments made with high-energy machines may help to elucidate the problem.

III. CHARGED UNSTABLE PARTICLES

§1. INTRODUCTION

The study of high-energy nuclear interactions using the counter-controlled cloud chamber and the electron-sensitive photographic emulsion has led to the discovery of a famiiy of charged unstable particles with masses greater than the well-known μ- and π-mesons. Four main empirical types have been identified, namely τ- and K-mesons, S-particles and charged V-particles. The τ-mesons have charge e, a mass of 977 m_e, and decay uniquely into three charged π-mesons. The K-mesons have charge e, masses in the range (900–1500) m_e, and decay into a light charged meson and one or more neutral particles. There are indications that the K-mesons are composite and include at least two types of meson, namely the κ-meson with a mass about 1100 m_e which decays into a μ-meson and two or more neutral particles, and the χ-meson with a mass within the range (900–1500) m_e, which decays into a charged π-meson of uniquely defined energy and one neutral particle. The main evidence for the τ- and K-mesons comes from experiments with the electron-sensitive emulsion.

S-particles have charge e and have been observed to decay at rest in thin lead plates inside a cloud chamber into a single light meson secondary. It is probable that many of the S-particles are the same as the χ-particles. Charged V-particles are particles of charge e observed in the cloud chamber, which decay into a single particle of charge e and one or more neutral particles. It is probable that many of the charged V-particles are identical with the K-particles and S-particles. Until the relationships between the four main types of decay process are established it is perhaps wise to retain the empirical classification. There is much evidence showing that K-mesons (and hence many of the charged V-particles) are produced directly in nuclear interactions.

Finally there is slight evidence for the existence of charged particles with masses about 500 m_e; provisionally these particles are known as charged ζ-particles.

§2. THE TAU MESON

2.1. *Introduction*

Since the discovery of the first τ-meson by Brown *et al.* (1949) eleven more examples have been found in the photographic emulsion and three in the cloud chamber. Details of seven τ-mesons are given in table 7, and recent examples are shown in Plates IX and XI. The data are entirely consistent with the original conclusion of Brown *et al.* that a particle of mass about 970 m_e decays spontaneously into three π-mesons. The photographic plate data give the most precise information about the masses of the primary and secondary particles and the mode of decay. The evidence is summarized in §§2.2 and 2.3.

2.2. *Primary Particles*

(i) Direct determination of the mass of the primary particle by various methods, for example, mean angle of scattering $\bar{\alpha}$ versus range R, or gap density (gap) versus range R (Powell 1950, Rochester and Rosser 1951), leads to a mean value for three good cases (B1, B2 and P6) of $(990 \pm 100)\, m_e$.

(ii) The change in grain density with distance along the track shows that the primary particle is slowing down as it approaches the point of decay (B1, B2, L3 and P6).

2.3. *Secondary Particles*

(i) The particles are coplanar in all cases to within the experimental errors of $\pm 1°$ or $\pm 2°$ (B1, B2, L3, L4, L5 and P6). It is therefore reasonable to assume that these particles are the only products of the decay. This is confirmed by the fact that momentum is conserved. A neutral secondary, if it exists at all, must be very light, and can carry away only a very small amount of energy.

Table 7. τ-Mesons

(1) Particle	(2) Medium	(3) Alt. (m) Absorber	(4) Length of prim. (μ)	(5) Obs. mass of prim. ($\times m_e$)	(6) Lengths of sec. (μ)	(7) Rel. masses of sec.	(8) K.E. of sec. (MeV)	(9) Total K.E. (MeV)	(10) Calc. mass of prim. ($\times m_e$)
Photographic-plate data									
B1	NT4	3450 7·5 cm Pb 1 cm brass	3100	990 ± 270 ($\bar{\alpha}R$) 1080 ± 160 (gap R)	45 2000 116	— 1 $1·02\pm0·10$	$1·04\pm0·10$ 31 ± 4 33 ± 4	65 ± 6	955 ± 12
B2	G5	3450 30 cm Pb	2070	1015 ± 280 ($\bar{\alpha}R$) 910 ± 220 (gap R)	6400 120 490	1 $0·88\pm0·10$ $1·03\pm0·05$	$19·0\pm0·4$ $24·2\pm2·0$ $32·0\pm2·0$	75 ± 4	975 ± 8
L3	G5	3450 3 m ice	740		420 1300 30	1 $1·05\pm0·20$ $0·96\pm0·25$	45 ± 12 $13·3\pm3$ $17·5\pm8$	76 ± 15	977 ± 29
L4	G5	3450 2·5 m ice	1900	1870 ± 500 ($\bar{\alpha}R$)			25 ± 5 24 ± 4 20 ± 4	69 ± 8	963 ± 16
L5	G5	3450 3 m ice			445 720 990	1 $0·93\pm0·12$ $1·09\pm0·17$	$33·5\pm5$ 31 ± 4 $9\pm1·5$	$73·5\pm7$	972 ± 14
P6	G5	4550 5 cm Al	8000	946 ± 140 ($\bar{\alpha}R$) 995 ± 150 (gap R)	2600 2000 100	1 0·99 0·95	$17·1\pm1·0$ $27·8\pm1·8$ $41·6\pm7·4$	$86·5\pm7·5$ mean K.E. $74·2\pm3·5$	998 ± 10 mean mass 973 ± 6
Cloud-chamber data									
Pa1	Argon	1750 Pb				~1 <2 ~1	155 ± 30* 350 ± 75* 210 ± 50*	~75	~975

* Measured momenta (MeV/c).

B1 Brown *et al.* (1949).
B2 Fowler *et al.* (1951).
L3 Harding (1950) and Herz *et al.* (1953).
L4 Harding (1950) and Herz *et al.* (1953).
L5 Hodgson (1951) and Herz *et al.* (1953).
P6 Ceccarelli *et al.* (1952).
Pa 1 Leighton and Wanlass (1952).

Mass of π-meson. Throughout this report we have assumed $m(\pi)=276\ m_e$, Wilson (1952), based on what is thought to be the most reliable method, viz. direct comparison with the mass of the proton by Birnbaum, Smith and Barkas. Barkas (1951) gives the error as $\pm 2\ m_e$. The measurements with the smallest stated errors are those of Barkas, Smith and Gardner, i.e. $m(\pi^+)=(277·4\pm1·1)\ m_e$ and $m(\pi^-)=(276·1\pm1·3)\ m_e$.

(ii) Some of the particles have been identified as π-mesons. One particle in B1 produced a two-prong star characteristic of a π^--meson. μ^--mesons occasionally produce two-prong stars, but the prongs are of very short lengths (George and Evans 1951). Again, the direct measurement of the mass of one of the secondaries of B2 by $\bar{\alpha}$ *vs.* gap density gave a value of $(285\pm20)\ m_e$.

(iii) Determination of the relative masses by the conservation of momentum and grain density shows that the masses are equal (B2, L3, L5 and P6: column

(7) of table 7). Taken with (ii) this suggests that all the secondary particles are π-mesons. Convincing evidence that this conclusion is correct, and that no combination of π- and μ-mesons is admissible, has been given by Ceccarelli *et al.* (1952) for P6. These authors first showed that the conservation of momentum was obeyed, and taking the measured value of the kinetic energy of one particle (a) on the assumption that it was either a π- or a μ-meson, they calculated the kinetic energies of the other two particles (b) and (c) again on the assumption that they could be either π- or μ-mesons. The theoretical values are compared with the experimental values in table 8. It is clear that the only cases in which there is good agreement between the theoretical and experimental values within the experimental errors are the first and last, that is, three π-mesons or three μ-mesons. However, for the μ-mesons, each of mass $210\,m_e$, the total kinetic energy is 63·8 MeV, resulting in a primary mass of $755\,m_e$, whereas for three π-mesons the total kinetic energy is 86·5 MeV leading to a primary mass of $998\,m_e$. The measured mass of the primary is $(970 \pm 100)\,m_e$, in good agreement with the assumption of decay to three π-mesons but well outside the value for three μ-mesons.

Table 8

Particle			Kinetic energy (MeV)				
a	b	c	E_a	E_b	E_c	E_b	E_c
				Theoretical		Experimental	
π	π	π	17·1	27·8	41·6	27·3	37·0
π	π	μ	17·1	27·8	49·7	27·3	28·9
π	μ	π	17·1	33·7	41·6	21·3	37·0
π	μ	μ	17·1	33·7	49·7	21·3	28·9
μ	π	π	13·6	17·8	27·0	27·3	37·0
μ	π	μ	13·6	17·8	33·5	27·3	28·9
μ	μ	π	13·6	22·5	27·0	21·3	37·0
μ	μ	μ	13·6	22·5	33·5	21·3	28·9

2.4. *Mass of the τ-Meson*

Having established the mode of decay of the τ-meson, the best value of the mass is found by adding the total kinetic energy (the Q-value) to the total mass of three π-mesons. It is seen from table 7 that the mean Q-value for these seven τ-mesons is $(74{\cdot}2 \pm 3{\cdot}5)$ MeV leading to a mass of $(973 \pm 6{\cdot}0)\,m_e$. In order to keep the data from different groups of workers consistent all the errors in table 7 are standard deviations. The weighted mean Q-value for the eleven τ-mesons known at the time of writing is 75·8 MeV leading to a mass of $977\,m_e$. The mass of the π-meson is taken as $(276 \pm 2)\,m_e$ after Wilson (1952). The lifetime and frequency of production of τ-mesons will be discussed later (III, §§ 8 and 9).

§ 3. K-MESONS

3.1. *Introduction*

Our knowledge of the τ-meson is precise and the mode of decay well established, but such is not the case for the other type of heavy meson. It appears from recent work that the charged mesons in the mass range $(900\text{–}1500)\,m_e$ are a composite group which includes two, and possibly more, types of particle

decaying in totally different ways. Apparently the lifetimes and frequencies of production of the particles are not very dissimilar, a result which makes the task of unravelling them very arduous.

3.2. *κ-Mesons and χ-Mesons*

The first κ-meson to be identified in the photographic emulsion was discovered by O'Ceallaigh (1951) who noticed while studying the decay electrons from μ-mesons that he had one example in which the momentum of the supposed electron was about four times greater than the maximum possible value. On measuring the mass of the primary particle he obtained a value of $(1320 \pm 170)\, m_e$. Shortly afterwards an event was observed which superficially appeared to be a $(\pi \rightarrow \mu)$ decay, but the range of the μ-meson was 1098 μ in contrast to the 600 μ for a $(\pi \rightarrow \mu)$ decay. Mass measurements on the supposed π-meson gave a value of

Table 9. K-Particles

	Primary particles				Secondary particles					
Event	Length (μ)	Mass ($\tilde{a}$ R) (m_e)	Mass (gap R) (m_e)	Mass (PR) (m_e)	Length (μ)	Grain density g^*	$p\beta$ (MeV/c)	p (MeV/c)	Mass (m_e)	Particle type
B1	4100	1260±260	1385±200	1518± 70	2200	0·97 ±0·04	250±35	280±40	<400	μ, π
B2	5670	1125±230	1125±150	(1100)	1100	—	11·8	—	210	μ
B3	535	(proton)	—	—	5900	0·96 ±0·026	157±14	180±15*	175±25	μ
B4	2100	1370±320	—	—	170	~1·0	—	—	—	—
B5	1540	1000±300	950±200	1050±150	8900	1·705±0·083	66±11	100±15*	220±15	μ
B6	2550	1000±240	(1050±200)	1450±150	150	~1·0	—	—	—	—
B7	380	(proton)	—	—	2500	1·090±0·050	174±29	210±35	270±45	π
B8	2550	1460±320	(900±200)	—	7650	1·140±0·025	192±18	230±20	330±30	π
B9	630	1380±600	—	—	19500	1·048±0·025	179±18	215±20	270±20	π
B10	1375	1100±330	—	—	275	~1·0	—	—	—	—
B11	540	1300±600	—	—	20	~1·0	—	—	—	—
P12	1300	1000+500 −350	—	—	20000	~1·0	174±24	210±30	<400	μ, π
M13	2080	1220±240	—	—	3440	0·94 ±0·05	150±27	—	—	μ
M14	5260	1030±120	1050±140	—	450	1·15 ±0·15	—	—	<400	—
G15	1885	1380±210	—	—	2085	1·14 ±0·05	108±18	—	190±30	μ

B1–B11 Menon and O'Ceallaigh (private communication). The photometric density measurements on particles B1, B2, B5 and B6 were made by von Friesen, Kristiansson, O'Ceallaigh and Walldeskog.
P12 Crussard, Mabboux *et al.* (1952).
M13 Levi-Setti and Lovati (private communication).
M14 and G15 Levi-Setti and Tomasini (1952).
Brackets indicate uncertain or preliminary measurements. The values of p for the secondary particles have been calculated on the assumption that they are π-mesons, except for B2, B3 and B5 where they are taken be μ-mesons.

$(1125 \pm 200)\, m_e$. This event is shown in Plate X. It was therefore suggested by O'Ceallaigh that there existed a new meson, the kappa meson, which decayed to a μ-meson and unobserved neutral particles. Since the original discovery more examples of decaying particles in the mass range $(900–1500)\, m_e$ have been found. The present position is summarized in table 9, in which are given fifteen accurately measured events in the photographic emulsion. The determination of masses has been made by the methods of mean angle of scattering *vs.* range ($\tilde{a}$, R), gap counting *vs.* range (gap, R), and by photometric track density *vs.* range (P, R), a method designed by von Friesen and Kristiansson (1950, 1952). The sources of the data are indicated at the foot of the table.

The main results may be summarized as follows: (i) Three secondary particles B2, B3 and B5 with kinetic energies of 6, 95 and 34 MeV respectively, have been

identified as μ-mesons. The secondary of B2 showed the typical $\mu \rightarrow e$ decay. (ii) Three secondary particles, B7, B8 and B9, have been identified as π-mesons. The evidence on which this identification is based and also the identification of the secondary of B3 as a μ-meson is shown diagrammatically in fig. 10. The secondary π-meson seems to have a unique kinetic energy of about 120 MeV.

Thus the observations point to the existence of two types of particle or two modes of decay of the same particle. On purely phenomenological grounds we prefer to regard the two modes of decay as representing two particles. The particle which decays to a μ-meson will be referred to as the kappa meson and the one which decays to a π-meson as the chi meson (χ). Since the μ-mesons from

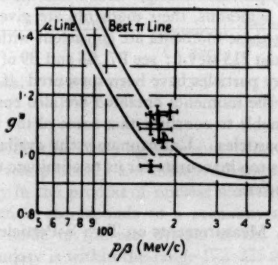


Fig. 10. Plot of g^* *vs.* $p\beta$ for secondaries of particles B3 (near μ-line), and B7, B8 and B9 (near π-line): the five points represent different parts of the tracks. See also table 7. (*After* Menon and O'Ceallaigh, private communication)

κ-mesons have very different energies the mode of decay of this particle must involve at least three particles, one charged and two neutral. The simplest scheme is then the following:

$$\kappa \rightarrow \mu + ?^0 + ?^0. \qquad \text{.....(III.1)}$$

Again, since the χ-meson seems to give a π-meson of unique energy the simplest form of the decay scheme is

$$\chi \rightarrow \pi + ?^0. \qquad \text{.....(III.2)}$$

There is some evidence that the κ and χ mesons have different masses, the κ-meson having a mass of about $1100\,m_e$ and the χ-meson a higher mass (see III, §5.4). Unfortunately, as can be seen from table 9, the results from two methods using the only χ-primary which is suitable for direct measurement are somewhat discordant, i.e. 1460 ± 320, and 900 ± 200, although in view of the large experimental errors they are not contradictory.

§4. S-PARTICLES

Bridge and Annis (1951) and Annis *et al.* (1952) have observed several examples of the decay of slow particles in plates in a multiplate cloud chamber (see table 1 for details). A typical event is shown in Plate XII, in which a particle associated with a penetrating shower passes through several plates and finally stops in a plate giving rise to a secondary penetrating particle. Annis *et al.* refer to this type of particle as a stopped or S-particle. Recently Bridge *et al.*

(1953 b) have completed the analysis of eight S-particles. Mass measurements by the method of scattering and range (Annis *et al.* 1953) give the mean mass of all the eight primaries as $(1460^{+410}_{-360})m_e$. The masses of the secondaries cannot be measured accurately because the tracks are at minimum ionization. The observed penetration of these secondary particles, in some cases as much as 40 g cm^{-2} of lead, proves that they are not electrons, and the values of $p\beta$ (measured by the method of scattering) combined with the ionizations indicate that they cannot be much heavier than the π-meson. Two of the secondary particles stop inside the chamber and their ranges can be measured; the data are given in table 10. The two ranges are almost identical, thus suggesting that the two secondary particles were produced with the same energy. Assuming the secondaries were π-mesons, their momenta are given in row 2 of table 10. It will be noticed that these momenta are consistent with the values obtained for the χ-particles (i.e. about 215 Mev/c: see B7, B8 and B9 of table 9). Several values of $p\beta$ for the secondary particles have been measured; if we assume the particles were π-mesons, then the momenta obtained are also consistent with 215 Mev/c. It is, therefore, reasonable to conclude that some of the S-particles are probably identical with the χ-particles. Unfortunately the available data are too meagre and the measurements too inaccurate for us to eliminate the possibility that some S-particles are κ-particles.

Table 10. Measurements on Two S-Particle Secondaries

	(1)	(2)
Range (g cm^{-2})	65·6–67·7	64·4–74·3
Momentum (Mev/c) (assuming π-meson)	210–214	209–222

§5. CHARGED V-PARTICLES

5.1. *Introduction*

The term charged V-particle is applied to those particles, other than μ- and π-mesons, which decay in flight in the cloud chamber to give a typical wide-angle forked track (Plate XIII). Since the initial discovery by the authors (Rochester and Butler 1947), these particles have been found in small numbers by Seriff *et al.* (1950), Armenteros *et al.* (1951, 1952), Bridge and Annis (1951), Fretter (1951), Thompson *et al.* (1951), Leighton *et al.* (1951), Astbury *et al.* (1952 a, b, 1953) and Bridge *et al.* (1953 b). It is generally agreed that charged V-particles are rarer than neutral V-particles, and that with the type of counter selection normally used (see II, §2) the ratio is about 1:6.

The data may conveniently be divided into two groups, slow and fast charged V-particles. In general the slow particles give more information than the fast ones, but they are much rarer. The methods of mass determination are those already described in II, §4.

5.2. *Slow Charged V-Particles*

The data on two slow V-particles (Leighton and Wanlass 1952 and Astbury *et al.* 1952, 1953), for which values of the masses can be obtained, are given in table 11. One of these particles is shown in Plate XIV.

In each case the authors have put forward arguments to show that the events are not due to the decay of V^0-particles coming upwards from interactions below

the chambers but are due to the decay of charged particles coming downwards. It will be noted that in one case the mass of the secondary particle is approximately equal to the mass of the μ- or π-meson. It is reasonable to conclude that these two decays are identical with those of the K-particles.

Bridge and Annis (1951) have obtained one example of the decay in the cloud chamber gas of a charged particle into a secondary which suffers nuclear scattering in an aluminium plate. This photograph is shown in Plate XV; the secondary particle, marked a, is seen to enter plate 4 and, after scattering, to emerge as particle b, and finally to stop in plate 9. From measurements of the scattering

Table 11. Data on Two Slow V-Particles

	Primary mass (m_e)	Ionization (× min)	Secondary momentum (MeV/c)	Mass (m_e)
Leighton and Wanlass	1200 ± 300	1·2–1·6	150 ± 15	250 ± 50
Astbury *et al.*	1100 ± 500	—	280 ± 50	—

and range, particle b is identified as a light meson. Since the particle suffered a nuclear scattering it was probably a π-meson. Assuming that the π-meson does not lose any energy in the process of nuclear scattering, the residual range at the point of production corresponds to a momentum of 145 MeV/c. If a transformation is now made to the rest system of the unstable particle, the momentum of the secondary is within the range 172–225 MeV/c. Providing that the energy loss at the scattering is small, the unstable particle is probably a χ-meson.

5.3. *Fast Charged V-Particles*

The results of Armenteros *et al.* (1952) will be taken as typical of the work on particles in this class. These authors, using the selection system and cloud chamber described in II, § 2, found 45 deflections in the gas of the chamber in 25 000 photographs of penetrating showers. Of the 45 deflections, 24 were excluded because of possible confusion with ($\pi \rightarrow \mu$) decays and large-angle scatterings. Six events out of 21 finally accepted could be measured completely, and direct mass measurements gave values of less than 1600 m_e for the primaries and 500 m_e for the secondaries. These results confirm those from slow V-particles, and also show that many of the particles are not the charged counterparts of the V_1^0-particle, whose mass is greater than the mass of the proton. In many cases the secondary particles have been observed to penetrate thick metal plates (see Plate XIII), and in no case is there evidence of interaction. This shows that these secondary particles could not have been electrons, but they could have been μ-mesons or π-mesons.

Armenteros *et al.* then examined the question of whether the fast charged V-decays could be described in terms of a simple two-body decay by comparing the observed distribution of transverse momentum p_t with the theoretical distribution given by equation (II.5), II, § 5. The results are shown in fig. 11, where the maximum value of p_t, i.e. p^*, has been taken to be the maximum observed value, that is 300 MeV/c. It is seen that the experimental distribution does not fit closely that for a unique two-body scheme. We are therefore forced to adopt either a three-body scheme or a mixture of decay schemes.

Finally, it may be noted that although some of the decays have taken place above lead plates, there is no evidence for the production of showers in the plates such as might have been expected if γ-rays were produced directly in the decay, or indirectly via the emission and decay of π^0-mesons.

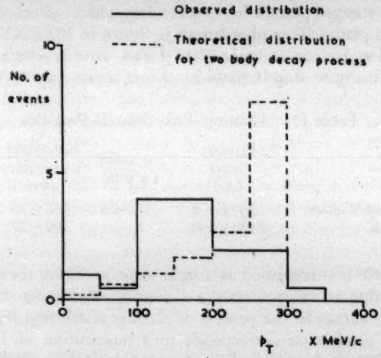


Fig. 11. The distribution of the transverse momentum components ($p^* = 300$ Mev/c for theoretical distribution). (*After* Armenteros *et al.* 1952; *reproduced by courtesy of the 'Philosophical Magazine'*)

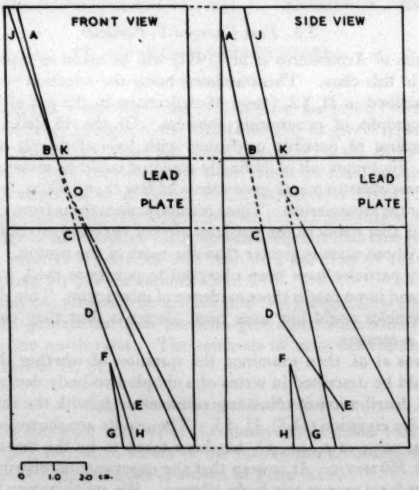


Fig. 12. The double decay.
(*After* Armenteros *et al.* 1952; *reproduced by courtesy of the 'Philosophical Magazine'*)

5.4. *The Double Decay*

One of the photographs of Armenteros *et al.* (1952), which is reproduced in Plate XVI and is so far unique, apparently shows a double decay. Two orthogonl projections of the event are shown in fig. 12. The negative particle represented by the track AB entered the plate and emerged as the track CD. The particle decayed at D, producing a negative secondary particle of momentum (178 ± 70) Mev/c at an angle of 24·5° with the direction of CD. The ionization of DE is less than twice minimum, and thus the mass is of the order or less than $(450 \pm 100)\, m_e$. About 1·25 cm from D, and immediately below, is a V^0-track. The point D lies in the plane of the fork to within the experimental error. A second particle interacted at O in the lead plate and produced two secondary particles. The plane of the V^0 also contains the point O within the rather large experimental errors of reprojection. It is therefore possible that the V^0 was produced at the

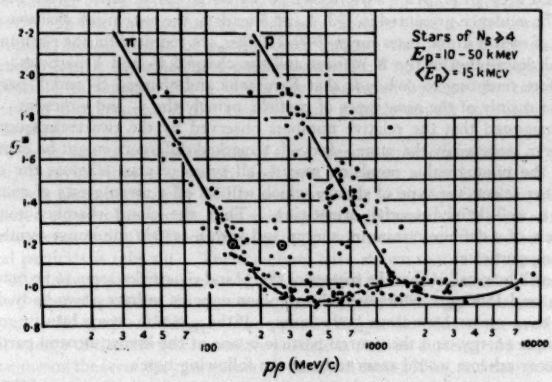


Fig. 13. g^* *vs.* $p\beta$. (*After* Daniel and Perkins, private communication)

point O, and appears by chance to be connected with D. If the coincidence is not fortuitous, the decay could have taken place according to the following simple scheme:

$$V^- \rightarrow V_1{}^0{}_{\text{or }2} + \text{light meson.} \qquad \ldots\ldots\text{(III.3)}$$

It has already been shown that there is, as yet, no conclusive evidence for a charged V-particle of mass greater than the proton. Thus it is probable that the V^0 particle is a $V_2{}^0$-particle. Armenteros *et al.* consider that the V^--particle could be identical with the χ-meson.

§6. CHARGED ZETA PARTICLES

Evidence for the existence of charged unstable particles in the mass range $(450–750)\, m_e$ seems to be gradually accumulating from both emulsion and cloud chamber work on high-energy interactions. In fig. 13 is shown a plot of g^*, the grain density of a track in terms of the density for particles of velocity almost equal to c, against $p\beta$, for the fast particles from nuclear interactions of mean energy 15 kMev, lying in the range 10–50 kMev. It is seen that while most of the particles

are grouped around the π-meson and proton masses, there are particles of apparent mass about $500\,m_e$.

The tracks of these particles have recently been carefully remeasured by Daniel and Perkins (private communication) and all except one are now consistent with a normal mass distribution centred on the π-meson mass. The exception is a particle which has a track length 11 mm in the emulsion and an ionization of 1·14 times minimum, leading to a mass of $(525 \pm 40)\,m_e$.

Leighton and Wanlass (1952) have reported three particles with masses in the range $(450–750)\,m_e$, the measurements being made by the method described in II, §4. These workers also claim to have found similar particles among the products of the decays of V^0-particles. Other groups of workers, employing similar methods, have yet to confirm these results.

§7. DECAY SCHEMES FOR K-MESONS AND CHARGED V-PARTICLES

The evidence presented in §§3, 4 and 5 leads to the conclusion that two main types of meson in the mass range $(900–1500)\,m_e$ are required for the phenomenological description of the K-mesons and the charged S- and V-particles. It is therefore tempting to conclude that K-mesons and charged S- and V-particles consist mainly of the *same* types of particle, namely the χ- and κ-mesons. It is not suggested that the relative numbers observed by the two techniques are, however, necessarily the same. Indeed, a marked difference might be expected since the photographic emulsion records all types of star, whereas the cloud chamber selects the type of shower which will set off a given group of counters with a well-defined spatial orientation. Thus the cloud chamber requires showers of a definite minimum energy and with a certain minimum number of ionizing particles.

It will be recalled that the masses of the χ- and κ-particles seem to be between $900\,m_e$ and $1500\,m_e$. Moreover, the χ-meson appears to have a two-body decay while the κ-meson has a three-body decay. If the χ-meson decays into a π-meson of unique energy, and the neutral particle is one of the known neutral particles, the decay scheme would seem to be of the following type:

$$\chi^{\pm} \rightarrow \pi^{\pm} + (\gamma, \nu, \pi^0, \text{ or } V_2^0). \qquad \text{.....(III.4)}$$

Taking the observed energy of the π-meson as 120 MeV, the masses of the χ-particles are $920\,m_e$ for γ and ν, $970\,m_e$ for π^0, and $1500\,m_e$ for V_2^0 of $900\,m_e$. In view of the fact that a decay involving a V^0-particle (probably V_2^0) has been observed in the cloud chamber, the decay scheme might be the following:

$$\chi^{\pm} \rightarrow \pi^{\pm} + V_2^0. \qquad \text{.....(III.5)}$$

If this scheme is true, then, as Powell (1952) has pointed out, it should be possible to find the V_2^0 by its characteristic decay in the photographic emulsion, for in this medium the χ-decay is at rest, and the direction of the neutral particle should be well defined. A careful search along strips of emulsion lying in the direction of the neutral particles defined by (III.5) for the three cases of χ-mesons found at Bristol has been made by Menon and O'Ceallaigh, but no V_2^0-decay, or any other evidence for a neutral particle, has yet been found. Because of the relatively short lengths of emulsion available in relation to the lifetime of the V_2^0, which is approximately 2×10^{-10} sec, this evidence cannot be taken to disprove the scheme (III.5). Bridge *et al.* (1953 b) have also searched their S-particle photographs for V_2^0-tracks or other evidence of the decay of the neutral

secondary particle since they conclude that S-particles are identical with χ-mesons. About 50 cm of path for the neutral particle have been examined but no decay products have been found. Bridge *et al.* conclude that if the neutral secondary is an unstable particle then its lifetime is probably longer than 5×10^{-9} second. Thus it is very unlikely that more than a small fraction of the S-particles and K-mesons decay according to scheme (III.5). Furthermore, no electron cascades were found which could have been produced by γ-rays, photons or the decay products of π^0-mesons coming from S-particles. We conclude that a definite decay scheme of the χ-meson cannot yet be assigned.

Very few accurate measurements have been made on slow charged V-decays. It is interesting now to consider the fast charged V-particles. It was shown in III, §5.3 that the preliminary measurements require a three-body decay scheme or a mixture of schemes. If χ-particles occur among the charged V-particles, then a number of the p_t-values should be close to p^* for the χ-particle, namely 215 MeV/c. In fact quite a number of values of p_t occur close to this value. The preliminary measurements of p_t, however, indicate that not all the charged V-particles can be explained as χ-particles (see III, θ5.3). Unfortunately it is not possible to separate the χ- from the κ-particles because the transverse momentum distributions of the two types of particle appear to overlap very considerably.

We come now to a consideration of the κ-meson, a particle of mass of about $1100\,m_e$ decaying into a μ-meson and two or more neutral particles. For simplicity a three-body decay will be assumed, but before specific schemes are considered a general argument, first used by Kofoed-Hansen (1951) and adapted by Armenteros *et al.* (1952), concerning the possible total maximum mass of the neutral particles is relevant. These authors have shown that in the three-body decay of a primary particle of mass M into a charged secondary particle of mass m_1 and two neutral particles of masses m_2 and m_3, the following relation holds: $(m_2+m_3)^2 = M^2 + m_1^2 - 2ME_{max}$. The quantity E_{max} is the value of the maximum kinetic energy of the charged secondary particle in the rest system. In the decay of the κ-meson the secondary particle is a μ-meson of mass $210\,m_e$, and the combined mass of the neutral particles can be calculated if the mass of the primary particle is assumed and also the value of E_{max}. Now E_{max} can be found easily from the value of p^*_{max}, which may reasonably be assumed to be not greater than the largest value of p_t observed in the cloud chamber photographs of charged V-particles or p^* in the photographic emulsion. The largest value of p_t or p^* observed is about 300 MeV/c, which leads to values of $m_2+m_3 \leqslant 250\,m_e$ if $M = 1200\,m_e$. Some typical results are shown in fig. 14. It is concluded that if the mass of the κ-meson is not greater than $1200\,m_e$, the total mass of the secondary neutral particles must be of the order of or less than the mass of the π-meson. Thus, if the neutral particles are known, the possible decay schemes would seem to be the following (Menon, private communication):

$$\left.\begin{aligned} \kappa^{\pm} &\rightarrow \mu^{\pm} + \nu + \nu \quad (a) \\ \kappa^{\pm} &\rightarrow \mu^{\pm} + \gamma + \nu \quad (b) \\ \kappa^{\pm} &\rightarrow \mu^{\pm} + \gamma + \gamma \quad (c) \end{aligned}\right\} \qquad \text{.....(III.6)}$$

$$\left.\begin{aligned} \kappa^{\pm} &\rightarrow \mu^{\pm} + \pi^0 + \nu \quad (a) \\ \kappa^{\pm} &\rightarrow \mu^{\pm} + \pi^0 + \gamma \quad (b) \end{aligned}\right\} \qquad \text{.....(III.7)}$$

The schemes have been set out in two groups in order to emphasize that the first group requires a κ-meson of lower mass than the second. It would seem that the following facts should be noted in attempting to choose the best scheme:

(i) Kappa mesons are probably produced directly in nuclear collisions (see §9). It follows that if they are not created in pairs, nor are the decay products of heavier particles, they must have integral spin.

(ii) There is no evidence for the presence of 'soft' secondaries (photons, electrons or π^0-mesons) among the decay products of charged V- or S-particles.

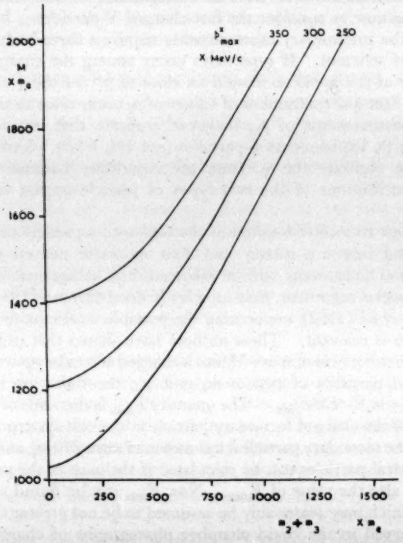


Fig. 14. Relationships between M and (m_2+m_3) for $m_1=210\ m_e$.
(*After* Armenteros *et al.* 1952; *reproduced by courtesy of the 'Philosophical Magazine'*)

If κ-mesons have integral spin, the five schemes are reduced to two, namely III.6 (*b*) and III.7 (*a*), and in view of the cloud chamber evidence, the first of these schemes is to be preferred to the second. As Menon has pointed out, it is possible that the photon spectrum from III.6 (*b*) is of such low energy that it would not have been observed in the cloud chamber. It seems unlikely that the electron cascades from the decay of the π^0-meson in scheme III.7 (*a*) could have been missed. Thus the best scheme might be

$$\kappa^{\pm} \rightarrow \mu^{\pm} + \gamma + \nu. \qquad \text{.....(III.8)}$$

A more detailed study of this problem will be needed before the true decay scheme is established.

§8. THE INTERACTION OF K-MESONS WITH MATTER

If K-mesons are nuclear-force mesons and are directly produced in high-energy interactions (see §9), it is to be expected that they will interact with matter. The general features of slow K-meson stars have been considered by Menon (private communication), who has shown that at least four different types could occur. A number of events have been reported (see for example Crussard *et al.* 1952, Fry and Lord 1952) which appear at first sight to represent the nuclear interaction of heavy mesons brought to rest. In most cases, however, it is not possible to show that the particle assumed to be the heavy meson is approaching the event, and in no case has a direct estimate of the mass of the heavy meson been made. The evidence is therefore indecisive. The results of Friedlander *et al.* (1953) show that if positive and negative heavy mesons exist in equal numbers, and if the decays observed are due to positive particles, then more than 75% of the negative particles do not produce observable disintegrations.

Concerning fast K-particles, there is some indication of interaction but no clear proof. Daniel *et al.* (1952) have found five interactions in 129·3 cm of track of the fast particles from 'jets' of energy greater than 50 kMev. The interaction length is therefore 26 ± 12 cm compared with 25 cm, the value expected if the interaction occurs according to the geometrical cross section. These workers have shown that 50% of the fast particles from this type of 'jet' are K-particles and 50% are π-mesons. It therefore seems plausible to infer that in this energy range both fast K-particles and π-mesons interact with the geometrical cross section.

§9. PRODUCTION OF CHARGED HEAVY MESONS

9.1. *Introduction*

The problem of the production of the different types of heavy meson—B-mesons and charged S- and V-particles—is one of fundamental importance not only for the understanding of their possible connection with nuclear forces, but also for their ultimate production by artificial means. The evidence available at present is meagre, but there are indications that modes of attack are being developed which should eventually yield many data. Because of the sparseness of the data the conclusions reached in this section will necessarily be tentative.

The difficulties encountered in treating the observational data are many, even if it is assumed that the only types of heavy meson are the τ-, κ- and χ-mesons. One fundamental difficulty lies in the fact that, with the present precision in mass measurement, the different types can only be properly recognized by the characteristic form of their decay. It follows that we have, at present, no exact knowledge of the relative numbers and types of those particles which do not decay. Again, even when the decay is observed, it is, in general, not possible to see the origin of the particle, and thus there is no certain evidence that the decaying particle was produced directly and did not come via the decay of a parent particle. It is possible that with the accumulation of data the difficulty will be overcome; indeed, as this report goes to press, news has come of the discovery of two slow K-mesons (particles M14 and G15, table 9) observed to be emitted directly from stars. This important discovery may be taken as strong evidence that K-mesons are either produced directly or are the products of parents with very short (i.e. $<10^{-14}$ sec) lifetimes.

A further difficulty in assessing the rate of production of the different types of particle is that every given experimental arrangement has its own particular efficiency of recording which depends upon the geometrical disposition of the recording apparatus, the lifetimes of the particles and many other factors (see Herz 1951). Since the essential data are only approximately known, it is not possible to get accurate values of the rates of production. It is therefore wise to put the main emphasis on the experiments in which the direct production of B-particles has been observed (Daniel *et al.* 1952, Daniel and Perkins, private communication). The results of these workers will be described in detail.

9.2. *The Creation of Heavy Mesons in 'Jets'*

The work of Daniel *et al.* (1952) is an extension and refinement of the Fowler method of the mass determination of shower particles by the measurement of coulomb scattering and grain density. Until recently it has been assumed that the grain density along the track of a singly charged particle in a photographic

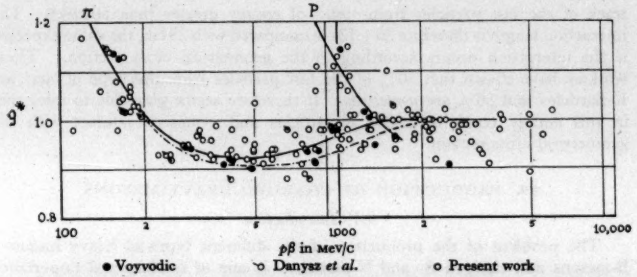


Fig. 15. Relation between grain-density, g^*, and momentum × velocity, $p\beta$, for the tracks of particles from 'stars' as determined by different observers. The full lines for π-particles and protons represent the best fit to the observation. The broken line represents a theoretical relation calculated by Voyvodic.

(*After* Daniel *et al.* 1952; *reproduced by courtesy of the 'Philosophical Magazine'*)

emulsion decreases with increasing energy until 3·5 times the rest energy (corresponding to a velocity of 0·94 **c**), and then remains substantially constant. Voyvodic (1952) has shown, however, that the grain density does not remain constant, but gradually increases until it reaches a constant value (plateau) of about 8% above minimum at ten times the rest energy. The actual magnitude of the observed increase depends upon the practical definition of a 'grain'. Some recent data are shown in fig. 15. It follows that care must be taken in defining the grain density of a track, and in the work under consideration the grain density g^* is always given relative to the grain density of an ultra-relativistic particle, i.e. $p\beta c > 10\mu c^2$ where μ is the mass of the particle. Whenever possible only those stars are selected with charged primary particles which fulfil this criterion and which have a track length greater than 4 mm. The values of the grain densities of the shower particles of the star are then normalized to the grain density of the track of the primary particle. If the primary track is not long enough the tracks of other fast particles in the same region of the plate are used. Great

care is taken to allow for change of grain density with depth, for distortion, for inaccuracies in the stage of the measuring microscope and for the effect of cell length and particle velocity on the scattering constant. Daniel and Perkins examined stars for which $N_h \geqslant 5$ and $n_s \leqslant 12$, produced by protons in the energy range 5–50 KMEV, with a mean energy, $\bar{E}_p$, of about 15 KMEV. The symbol N_h means the number of heavily ionizing tracks (i.e. $g^* > 1{\cdot}5 g_{\text{plateau}}$), and n_s the number of shower particles.

Some typical results are shown in fig. 13, from which it will be seen that the shower particles are almost all π-mesons and the other emitted particles are protons and B-particles. The (proton/π-meson) mass ratio is found to be $6{\cdot}6 \pm 0{\cdot}25$, in excellent agreement with the value calculated from the known masses of the proton and the π-meson. Daniel *et al.* (1952) have also obtained

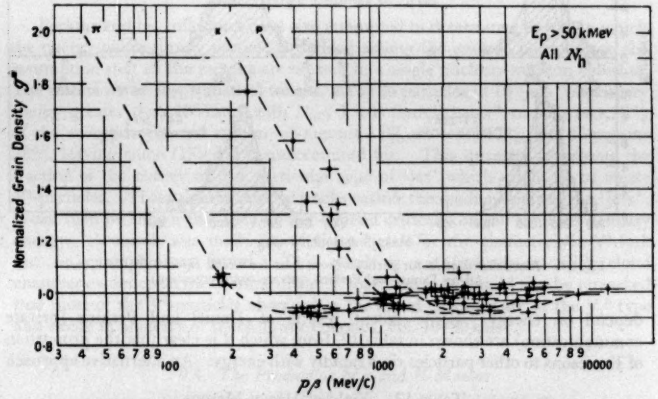


Fig. 16. Measurements on 'stars' with $E_p > 50$ KMEV for all values of N_h.
(*After* Daniel *et al.* 1952; *reproduced by courtesy of the 'Philosophical Magazine'*)

similar results on high-energy stars with few heavily ionizing tracks, the so-called 'jets', because it was inferred from the work of Camerini *et al.* (1951) that these would provide the most favourable place in which to look for B-particles. Examples of jets in which the primary particles have energies of 3000 KMEV, 1000 KMEV and 200 KMEV are shown in Plates XVII and XVIII. A typical set of (g^*, $p\beta$) curves for $E_p > 50$ KMEV, ($\bar{E}_p \sim 300$ KMEV) is shown in fig. 16. It will be observed that there is a marked decrease in the number of slow π-mesons and protons and a significant increase in the numbers of particles with masses about $1200\, m_e$. The mass distribution based on curves similar to fig. 16, but including lower energies, is shown in fig. 17, from which it will be seen that particles also occur in the range $(900–1400)\, m_e$. Since the distribution is grouped about $1200\, m_e$, and does not extend much below $1000\, m_e$, it is assumed that these particles are mainly K-mesons and that very few are τ-mesons. Other evidence in support of this conclusion will be presented in § 9.4. The observed numbers of K-mesons which come to rest in the emulsion are in agreement with this

assumption. It is presumed that the observed K-mesons are a mixture of κ- and χ-mesons and are also representative of charged V-particles. There is fairly good evidence for the direct production of *both* κ- and χ-mesons.

Even apart from the question of direct production, the observations give valuable information on how the energy of the primary particle is distributed among the emitted particles, and how the numbers of the new particles

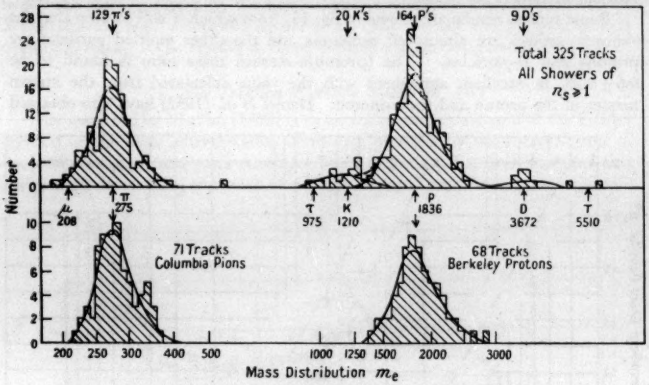


Fig. 17. Mass distribution for particles from stars for which E_p=10–50 kMev, $n_s \geq 1$. (*After* Daniel and Perkins, private communication)

depend on energy. Some recent results of Daniel and Perkins (private communication) are shown in table 12, from which it is clear that the proportion of K-mesons to other particles rises rapidly with energy. An alternative approach

Table 12. Yield of Heavy Mesons

Star type	N_K/N_π β=0·5–0·84	N_K/N_π p=330–950 Mev/c	(Identified K shower particles) (All shower particles) ($g^* < 1{\cdot}4$)	E_K/E_π p=330–950 Mev/c	Average energy E_p (kMev)
n_s=1, 2, 3	0·13 ± 0·05	0·09 ± 0·04	0·015 ± 0·006	0·17 ± 0·06	5 (1·5–8)
$n_s \geq 4$	0·28 ± 0·08	0·20 ± 0·06	0·018 ± 0·004	0·36 ± 0·1	20 (6–40)
'Jets' (E_p > 50 kMev)	—	0·5 ± 0·2	~0·08	1·0 ± 0·3	200 (50–3000)

to the problem has been made by Daniel *et al.* (1952). The method consists in comparing the ratio of π^0-mesons (which give rise to electron pairs close to the star) to charged mesons, i.e. $R_1 = N(\pi^0)/N(\pi^\pm)$ in relatively low-energy stars and in 'jets'. For the lower-energy range ($\bar{E}_p \sim 25$ kMev, $n_s > 6$), $R_1 = 0{\cdot}56 \pm 0{\cdot}12$. For 'jets', however, the ratio is $0{\cdot}33 \pm 0{\cdot}07$. The new ratio may be written

$$R_2 = N(\pi^0)/[N(\pi^\pm) + N(\mathrm{K}^\pm) + N(\mathrm{P})].$$

Allowing for the small number of protons, and making other small corrections, it is found that

$$R_3 = \frac{N(\pi^0)}{N(\pi^\pm) + N(\mathrm{K}^\pm)} = 0{\cdot}35 \pm 0{\cdot}07.$$

Whence, from the experimentally observed value of R_1, $N(K^{\pm})/N(\pi^{\pm}) = 0{\cdot}6 \pm 0{\cdot}5$, in good accord with the value given in table 12, although the figures in the table refer to only a narrow energy range for both K- and π-mesons. Lal *et al.* (1952 a, b) have recently given a value for R_2 of $0{\cdot}86 \pm 0{\cdot}1$ for the secondary particles of a 'jet' produced by a magnesium nucleus of energy 2×10^{14} ev. This high value for R_2 seems to be due to inadequate correction for 'soft' cascade development. A lower value for R_2 is consistent with the results of the Rochester group and the analysis of a high-energy event found by Mulvey at Bristol. This even was produced by a photon of energy 3×10^{13} ev and gives a value for R_2 of $0{\cdot}24 \pm 0{\cdot}09$. Observations on the development of the 'soft' cascade show that R_2 lies within the range 0·18–0·40.

9.3. *Energy Balance in 'Jets'*

Perkins and his colleagues have also attempted to determine the way in which the energy of a primary particle is divided among the emitted particles, on the assumption that all the mesons are created in a single nucleon–nucleon collision. The results are of low statistical weight. For 26 particles in 10 'jets' of primary energy greater than 100 kMev with $N_h \leqslant 3$ it is found that $\pi^{\pm}$-mesons take 30% of the energy in the rest system, π^0-mesons 15%, (P + *M*) 12%, and K-mesons 28%, leaving some (15 ± 12)% unaccounted for. This quantity represents the fraction of the energy of this particular type of 'jet' which might go to create V^0-particles. These figures refer to tracks *outside* the median angle of the 'jets'; it can then be shown that they do not depend critically on the assumed primary energy. A search was made round these 'jets' for the characteristic V^0-fork but no example was found. This result is not inconsistent with cloud chamber evidence on the lifetime of the V^0-particles. It should also be remarked that most of the V^0-particles observed in the cloud chamber are of the $V_1{}^0$ type and occur in showers of much lower energies, viz. 10–50 kMev.

9.4. *The Production of τ-and K-Mesons*

The evidence for production which may be deduced from slow τ-mesons and K-mesons may be summarized as follows:

(i) Slow K-mesons are created directly in high energy interactions (Bridge and Annis 1951, Levi-Setti and Tomasini 1952 and Isachsen *et al.* 1953).

(ii) Slow τ-mesons are created directly in high energy interactions (Ceccarelli *et al.* 1953). There is some evidence for the preferential production of τ-mesons in materials rich in hydrogen, like ice (Herz *et al.* 1953). Thus the Bristol group found $(0{\cdot}0003^{+0{\cdot}0004}_{-0{\cdot}0002})$ τ-mesons per slow π-meson under various absorbers, while Hodgson observed $(0{\cdot}004^{+0{\cdot}004}_{-0{\cdot}002})$ τ-mesons per slow π-meson in ice. Hodgson suggested that the reason for this difference might be the strong self-absorption of the τ-meson or its neutral parent in nuclear matter.

9.5. *The Production of Charged V-Particles*

The main facts on the production of charged V-particles which seem to be indicated by the most recent work are the following:

(i) The observed rate of charged V-particles is about 15% of the rate of neutral V-particles in the energy range 10–30 kMev. There is one charged V-decay to some 600 shower particles (Barker *et al.* 1952).

(ii) The observed rate in cloud chambers is consistent with the results of Daniel and Perkins (private communication) on the assumption of a reasonable value for the lifetime (Herz 1951, Barker *et al.* 1952).

(iii) Charged V-particles are sometimes produced in relatively low-energy interactions, i.e. of the order of 5 kMev (Astbury *et al.* 1952, 1953, Barker, private communication).

The general conclusion which may be drawn from these results is that charged V-particles are very similar to, and are probably identical with, K-mesons. The fact that these particles can be produced in low-energy interactions may have important consequences for their artificial production.

§ 10. LIFETIMES OF HEAVY MESONS

10.1. *Introduction*

Only very approximate estimates of the lifetimes of charged heavy mesons have been made so far, but the general conclusion from all the experimental evidence available is that the mean lifetimes lie in the time interval $(10^{-10} - 10^{-8})$ sec (Fowler *et al.* 1951, Barker *et al.* 1952, Astbury *et al.* 1952, 1953). As more slow mesons are found, direct timing methods may become possible even with the main experimental techniques now in use. Precise values for the different types of particle may not be obtained until the particles are produced by artificial means.

We proceed now to describe recent work on the τ-meson by Herz *et al.* (1953) and on charged V-particles by Barket *et al.* (1952) and Astbury *et al.* (1952, 1953). The results for charged V-particles represent an average lifetime of an unknown mixture of κ- and χ-mesons.

10.2. *The τ-Meson*

The estimates of the lifetime of this particle were made as a result of the observation of three slow τ-mesons in photographic emulsions exposed in an iron pipe at depths of from 0–9 m in a glacier. Herz *et al.* (1953), following methods developed by Herz (1951), have calculated, as a function of the assumed lifetime, the numbers of τ-mesons present just outside the pipe for each meson decaying in the emulsion, (*a*) at rest, and (*b*) in flight. Taking a reasonable value for the kinetic energy of the meson at formation, it is found that the numbers for (*a*) are 1·07, 1·93, 713 and 3×10^{28} for lifetimes of 10^{-7}, 10^{-8}, 10^{-9} and 10^{-10} sec respectively. If the τ-meson interacts with the geometrical cross section the numbers are somewhat larger but the absolute magnitudes remain the same. Now it is known that only very small numbers of τ-particles are produced, and Herz *et al.* consider that the only acceptable frequencies are those for lifetimes longer than 10^{-9} second. The calculated values for (*b*) are 0·3, 250 and $2{\cdot}2 \times 10^{29}$ for lifetimes of 10^{-8}, 10^{-9} and 10^{-10} sec respectively. A detailed examination of 0·27 cm^3 of emulsion under high magnification failed to reveal any examples of decay in flight, and Herz *et al.* estimate that there could have been no more than two such events per slow τ-meson. Thus in this case also the results lead to a value for the lifetime longer than 10^{-9} second. The two examples of decay of τ-mesons in the cloud chamber found by Leighton and Wanlass (1952) are consistent with this estimate.

10.3 *Charged V-Particles*

The results of Barker *et al.* (1952) are based on the observation of 22 decays in a circular chamber, diameter 28 cm, and those of Astbury *et al.* (1952a) on 13 decays in a square chamber of side 54 cm. The essential data that are needed to make a determination of the lifetime are the proper time intervals t and T, i.e. the times measured in the rest system of the particle between entry and the point of decay, t, and between entry and the point of departure had the decay not occurred, T. Once the set of traversal times t and T are obtained the most probable lifetime can be deduced following the procedure of Peierls (1935), Hole (1947) and Bartlett (1953). It follows from this treatment that if the average value of t/T is close to 0·5 the lifetime τ_0 is probably longer than the average value of T.

Wilson and Butler (1952) have discussed the reduction of the data obtained from cloud chamber photographs to the essential quantities t and T, and they have suggested that since separate groups of workers are collecting data slowly, they should be presented in a form which will allow the best value to be obtained. The observational data given by the photographs consist of the momentum of the primary charged V-particle p, the path length in the chamber before decay, and the potential path length, that is, the distance between the points of entry and departure of the particle. Wilson and Butler point out that these lengths should be measured inside fiducial surfaces located at some distance within the geometrical dimensions of the chamber, the actual positions being determined by factors such as the distribution of illumination in the chamber, the efficiency of condensation, etc. The symbols l and L will be used for the corrected path length and potential path length respectively. From these quantities, and the mass of the particle m, the proper times are found from the expressions

$$t = \frac{l}{c}\left(\frac{m}{p}\right) \quad \text{and} \quad T = \frac{L}{c}\left(\frac{m}{p}\right),$$

where the momentum and mass are measured in Rossi units. Barker *et al.* assume m to be $1200\,m_e$, and this value is also accepted by Astbury *et al.* except in two of their cases where direct estimates can be made. The results of Barker *et al.* give an arithmetic mean value of t of 1×10^{-10} sec and Astbury *et al.* of 6×10^{-10} sec. Peierls showed that this mean is a *minimum* value of the lifetime. It is concluded that the lifetime is unlikely to be less than 10^{-10} sec. An upper limit can be derived by an examination of how the numbers of decays observed in the cloud chamber vary with various assumed lifetimes, following a procedure which is rather similar to the one developed by Herz (1951). The relationship between the lifetime and frequency of production found by Barker *et al.* is summarized in fig. 18. Curves A and A′ show the ratio of the number of charged V-particles produced in the lead above the chamber N_V, to the number of shower particles N_s, as a function of lifetime for distances between the lead and the chamber of 5 cm and 10 cm respectively. Curve B shows the ratio of the number N of decays observed in the chamber, to N_V as a function of τ_0. This curve shows in an instructive way how efficiently this particular arrangement of absorber and chamber records charged V-particles of different lifetimes, Now Daniel *et al.* (1952) have shown that in penetrating showers of mean energy about 15 Mev K-mesons constitute less than 20% of all shower particles. The

Pic-du-Midi chamber probably selects showers of approximately this energy, and hence it is concluded that the mean lifetime of charged V-particles is unlikely to be much greater than 10^{-8} second. If the frequency of production were as low as 1% the lifetime would be of the order of 10^{-9} second.

Astbury *et al.* (1953) have recently examined the tracks of slow, non-decaying negative particles and they have found ten which were probably due to heavy mesons. The total observed time of flight of these particles was $1{\cdot}8 \times 10^{-8}$ sec, and after combining this time with the number of slow V-particles observed to decay in the cloud chamber it is concluded that the mean lifetime is probably greater than 4×10^{-9} second.

IV. SUMMARY OF THE DECAY SCHEMES OF NEUTRAL AND CHARGED UNSTABLE PARTICLES

The detailed properties of the neutral and charged unstable particles discussed in sections II and III are summarized in table 13 (see also Powell 1952, Heisenberg 1952). Particles and decay schemes for which there is no direct evidence are framed in. It appears that in the mass range (500–2200) m_e and the lifetime

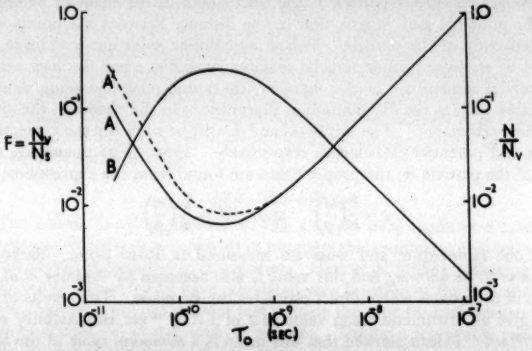


Fig. 18. The relations between the lifetime and frequency of production of $V^{\pm}$-particles.
(i) Curves A and A' show N_V/N_S as a function of τ_0 for distances between the absorber and the chamber of 5 and 10 cm respectively.
(ii) Curve B shows N/N_V as a function of τ_0 for a distance of 5 cm between the absorber and the chamber.
(*After* Barker *et al.* 1952; *reproduced by courtesy of the* '*Philosophical Magazine*')

range (10^{-8}–10^{-10}) sec there are at least five different types of unstable particle, of which three are charged, $\chi^{\pm}(V_{\chi}^{\pm}, S)$, $\kappa^{\pm}(V_{\kappa}^{\pm}, S)$ and $\tau^{\pm}$, and two are neutral, V_1^0 and V_2^0. Although the charged particles are assumed to be both positive and negative, definite empirical evidence that both signs occur is lacking in certain cases. It is also not certain that the particles included within each of the five main groups are of unique mass and type. Only in the case of the τ-meson and the V_1^0-particle can the decay schemes be considered established and, therefore, the masses uniquely determined.

Furthermore, it is quite possible that other types of particles occur in the stated mass range. Thus, as already pointed out, there are indications of charged and

neutral particles, the $\zeta^{\pm}$ and ζ^{0} particles, with masses about $500\,m_e$. Again a charged V_1-particle corresponding to the neutral V_1-particle may exist. Possible decay schemes for such a particle are given in the table.

The position regarding the masses and decay schemes of both charged and neutral particles in the mass range (800–1500) m_e is still confused. The charged mesons seem to be definitely of two types, one decaying to a π-meson of unique energy (the χ-meson) and another to a μ-meson which seems to have a wide

Table 13. V-Particles and Heavy Mesons
(December 1952)

Particle	Decay products	Mass (m_e)	Q (MeV)	Mean lifetime (sec)	Spin
$\chi^{\pm}$, $V^{\pm}$, $S^{\pm}$	$\pi^{\pm}+$ [neutral V_2^0?]	900–1500		10^{-8}–10^{-10}	—
$\kappa^{\pm}$, $V^{\pm}$, $S^{\pm}$	$\mu^{\pm}+$ [$\gamma+\nu$]	1100		10^{-8}–10^{-10}	Integral
$\tau^{\pm}$	$\pi^{\pm}+\pi^{+}+\pi^{-}$	977 ± 6	75·8	$>10^{-9}$	Integral
[$V^{\pm}\rightarrow\pi^{\pm}+\pi^{0}+\pi^{0}$]					Integral
V_2^0	$\pi^{\pm}+\pi^{-}$	950		$\sim2\times10^{-10}$	Integral
	$\pi^{+}+\pi^{-}+?^{0}$				
	$\pi^{\pm}+\tau^{\mp}$ or $\zeta^{\mp}$				
$\zeta^{\pm}$?	—	(500)			
ζ^{0}?	$\pi^{+}+\pi^{-}$	(500)			Integral
V_1^0	$p^{+}+\pi^{-}$	2190 ± 5	40 ± 3	3×10^{-10}	Half integral
	$n^{0}+\pi^{0}$	—	—	—	Half integral
[$V^{\pm}_1$]	$n+\pi^{\pm}$	—	—	—	Half integral
	$p^{+}+\pi^{0}$	—	—	—	Half integral

Particles and decay schemes framed in have not been definitely observed.

spread of energies (the κ-meson). Concerning neutral particles there is definite evidence for the existence of at least one type (the V_2^0), but the mode of decay is still uncertain. Many of the measured examples can be explained by the decay of a particle of mass about $950\,m_e$ into two charged π-mesons. It is possible, however, that there are alternative decay schemes, and also that a third body may be produced. Several interesting speculations can be made. For example, one type of V_2^0-particle may be the neutral particle which appears in the decay of the

χ-meson. Again, it is tempting to suggest that one type of V_2^0-particle may be the neutral counterpart of the τ-meson and decay according to the scheme

$$V_2^0(\tau^0) \rightarrow \pi^+ + \pi^- + \pi^0. \qquad \text{.....(IV.1)}$$

It is unlikely, however, that this scheme can explain all the V^0-tracks because many of the cases found by Armenteros *et al.* give a mass of about $950\,m_e$ considering only the two charged secondaries. Therefore the addition of the π^0-meson mass increases the masses of the examples to values considerably greater than the τ-meson mass of $977\,m_e$. In addition, there is no positive evidence for the presence of the π^0-meson.

Leighton *et al.* have reported one V_2^0-decay in which the mass of the negative secondary particle lies in the mass range $(450–1000)\,m_e$. Following Goldhaber, this observation suggests a scheme of the type

$$V_2^0 \rightarrow \pi^{\pm} + \tau^{\mp}. \qquad \text{.....(IV.2)}$$

For a Q-value of 60 Mev the mass of the V_2^0 is about $1400\,m_e$, in which case the V_2^0-particle could be regarded as the neutral counterpart of the χ-particle. The decay of the χ-particle could then be of the following type:

$$\chi^{\pm} \rightarrow \pi^{\pm} + \tau^0. \qquad \text{.....(IV.3)}$$

V. FUTURE OUTLOOK

The observation of a particular type of particle is undoubtedly closely related to the techniques which are available, and there seems to be no reason to doubt that as new techniques are developed new particles will be discovered. It seems unlikely therefore that we know yet the main types of particle which exist even apart from this possibility; the subject is already one of great complexity, and it will probably be many years before the underlying pattern is discovered. Indeed the present position bears some resemblance to the early days of spectroscopy and radioactivity.

While it is perhaps unwise to attempt to foreshadow future developments, there are one or two comments which may be made. It is clearly important, for example, to collect many more empirical data, of the type already obtained, to attempt to check the decay schemes which have been tentatively suggested. This might involve, in the case of the emulsion-technique, the use of 'stripped' emulsions and thicker emulsions exposed in balloons at low geomagnetic latitudes (Powell 1953), with distortion-free development and, in the case of the cloud chamber, new methods of selection. There are promising signs that such methods are now being developed, for example by Salvini and Reynolds (1951), Salvini and Kim (1951), Althaus (1951), Barker, Sard and Sowerby (1952), and Duerden and Hyams (1952). Furthermore, for both of the main techniques so far used there is need for greater precision in mass determination. In the emulsion, which is already superior to the cloud chamber in this respect, a new method is needed for the determination of masses which are relatively close together, as, for example, the μ- and π-mesons, or the κ- and χ-mesons. In the cloud chamber all the methods of mass determination commonly in use, whether by momentum *vs* ionization or momentum *vs* range, need re-examining. The problem is a difficult technical one, but on its solution depends the whole future of mass determinations using the cloud chamber. A technique which has properties in common with both the photographic emulsion and the cloud chamber is the high pressure chamber (Evans *et al.* 1951).

With the completion of the great machines which can produce particles of energy of a few kMeV it seems it will be only a matter of time before heavy mesons are artificially produced. Once these particles are available great progress should be made. The greater numbers of such particles should lead to precise determinations of the masses (perhaps by direct comparison with the mass of the proton) and lifetimes. Moreover, it may become possible to choose the type of meson to be studied by varying the energy of the bombarding particle.

Clearly, then, there is an immense task ahead, and it would seem that it will be many years before the various particles are known and their connection with the nuclear-force particles of physics established. It is therefore perhaps unwise to comment at this stage in detail on the schemes outlined in the table. One striking fact (Powell 1952) is that all particles except the κ-meson produce π-mesons of various sorts, and it is tempting to ask if most of the new particles are somehow made up either of π-mesons or of π-mesons and nucleons. Another question of interest is whether the V_1^0-particle, which decays to a nucleon, can be considered as a form of excited nucleon. If this is the case it may be possible to produce V_1^0-particles without first creating the nucleon (see Plates II and VII).

ACKNOWLEDGMENTS

It is a pleasure to acknowledge the helpful comments of our colleagues Mr. K. H. Barker and Dr. W. O. Lock, who kindly read this report in manuscript.

We are grateful to the many authors who sent us information or copies of their papers in advance of publication. Papers and photographs are acknowledged in the text. Dr. Menon and Dr. Perkins of Bristol University kindly checked some of the data given in section III, and were most helpful in every way.

Finally we wish to express our thanks for the permission granted by the various authors and publishing bodies for the reproduction of diagrams and photographs.

REFERENCES

General

Butler, C. C., 1951, *Progress in Cosmic Ray Physics* (Amsterdam : North-Holland Publishing Company), p. 105.

Heisenberg, W., 1952, Report of International Physics Conference, Copenhagen, p. 35.

Powell, C. F., 1950, *Rep. Prog. Phys.*, **13**, 350 (London : Physical Society); 1952, Report of International Physics Conference, Copenhagen, p. 29.

Rochester, G. D., and Rosser, W. G. V., 1951, *Rep. Prog. Phys.*, **14**, 227 (London : Physical Society).

Rochester, G. D., and Wilson, J. G., 1952, *Cloud Chamber Photographs of the Cosmic Radiation* (London : Pergamon Press Ltd.), p. 102.

Rossi, B., 1952, *High Energy Particles* (New York : Prentice-Hall Inc.), p. 207.

Neutral V-Particles

Althaus, E. J., 1951, *Phys. Rev.*, **81**, 647; *Ibid.*, **83**, 896.

Armenteros, R., Barker, K. H., Butler, C. C., and Cachon, A., 1951 b, *Phil. Mag.*, **42**, 1113.

Armenteros, R., Barker, K. H., Butler, C. C., Cachon, A., and Chapman, A. H., 1951 a, *Nature, Lond.*, **167**, 501.

Barker K., H., Sard, R. D., and Sowerby, M. G., 1952, *Phil. Mag.*, **44**, 46.

Bridge, H. S., and Annis, M., 1951, *Phys. Rev.*, **82**, 445.

Bridge, H. S., Courant, H., and Rossi, B., 1952, *Phys. Rev.*, **85**, 159.

Bridge, H. S., Peyrou, C., and Safford, R., 1953 a, *Phys. Rev.*, in the press.

Brueckner, K. A., and Thompson, R. W., 1952, *Phys. Rev.*, **87**, 390.

CARTER, R. S., and WHITTEMORE, W. L., 1952, *Phys. Rev.*, **87**, 494.
DEUTSCHMANN, M., 1952, *Z. Naturforsch.*, **7a**, 142.
FRETTER, W. B., MAY, M. M., and NAKADA, M. P., 1953, *Phys. Rev.*, **89**, 168.
GHOSH, S. K., JONES, G. M. D. B., and WILSON, J. G., 1952, *Proc. Phys. Soc.* A, **65**, 68.
HOPPER, V. D., and BISWAS, S., 1950, *Phys. Rev.*, **80**, 1099.
LEIGHTON, R. B., and ALFORD, W. L., 1952, *Bull. Amer. Phys. Soc.*, **27**, 18.
LEIGHTON, R. B., WANLASS, S. D., and ALFORD, W. L., 1951, *Phys. Rev.*, **83**, 843.
LEIGHTON, R. B., WANLASS, S. D., and ANDERSON, C. D., 1953, *Phys. Rev.*, **89**, 148.
ROCHESTER, G. D., and BUTLER, C. C., 1947, *Nature, Lond.*, **160**, 855.
SERIFF, A. J., LEIGHTON, R. B., HSIAO, C., COWAN, E. W., and ANDERSON, C. D., 1950, *Phys. Rev.*, **78**, 290.
THOMPSON, R. W., COHN, H. O., and FLUM, R. S., 1951, *Phys. Rev.*, **83**, 175.

ζ^0-Particles

BRUECKNER, K. A., and WATSON, K. M., 1952, *Phys. Rev.*, **87**, 621.
COSYNS, M. G. E., 1949, *Nuovo Cim.*, **6**, 397; 1950, Report Harwell Conference, p. 19.
DANIEL, R. R., DAVIES, J. H., MULVEY, J. H., and PERKINS, D. K., 1952, *Phil. Mag.*, **43**, 753.
DANYSZ, M., LOCK, W. O., and YEKUTIELI, G., 1952, *Nature, Lond.*, **169**, 364.
DANYSZ, M., HARRIS, G. G., JURITZ, J. W. F., and LOCK, W. O., 1953, *Phil. Mag.*, in the press.
KAPLON, M. F., and RITSON, D. M., 1952, *Phys. Rev.*, **87**, 664.
LAI, D., PAL, Y., PETERS, B., and SWAMI, M. S., 1952 a, *Phys. Rev.*, **87**, 545; 1952 b, *Proc. Ind. Acad. Sci.*, **36**, 75.
MEI, J. Y., and PICKUP, E., 1952, *Phys. Rev.*, **86**, 796.

τ-Mesons

BARKAS, W. H., 1951, Physics Distribution No. 1285, Radiation Laboratory, University of California, U.S.A.
BROWN, R. H., CAMERINI, U., FOWLER, P. H., MUIRHEAD, H., POWELL, C. F., and RITSON, D. M., 1949, *Nature, Lond.*, **163**, 82.
CECCARELLI, M., DALLAPORTA, N., MERLIN, M., QUARENI, G., and ZOM, G. T., 1953, *Nuovo Cim.*, in the press.
CECCARELLI, M., DALLAPORTA, N., MERLIN, M., and ROSTAGNI, A., 1952, *Nature, Lond.*, **170**, 454.
FOWLER, P. H., MENON, M. G. K., POWELL, C. F., and ROCHAT, O., 1951, *Phil. Mag.* **42**, 1040.
GEORGE, E. P., and EVANS, J., 1951, *Proc. Phys. Soc.* A, **64**, 193.
HARDING, J. B., 1950, *Phil. Mag.*, **41**, 405.
HERZ, A. J., 1951, Report of Conference on V-Particles and Heavy Mesons, Bristol, p. 25.
HERZ, A. J., HODGSON, P. E., and TENNENT, R. M., 1953, *Phil. Mag.*, **44**, 85.
HODGSON, P. E., 1951, *Phil. Mag.*, **42**, 1060.
LEIGHTON, R. B., and WANLASS, S. D., 1952, *Phys. Rev.*, **86**, 426.
WILSON, R., 1952, Report of International Physics Conference, Copenhagen, p. 42.

K-Mesons

CAMERINI, U., DAVIES, J. H., FOWLER, P. H., FRANZINETTI, C., MUIRHEAD, H., LOCK, W. O., PERKINS, D. H., and YEKUTIELI, G., 1951, *Phil. Mag.*, **42**, 1241.
CRUSSARD, J., LEPRINCE-RINGUET, L., MABBOUX, C., MORELLET, D., TREMBLEY, J., and ORKIN-LECOURTOIS, A., 1952, Report of International Physics Conference, Copenhegen, p. 31.
CRUSSARD, J., MABBOUX, C., MORELLET, D., TREMBLEY, J., and ORKIN-LECOURTOIS, A., 1952, *C. R. Acad. Sci., Paris*, **234**, 84.
DANIEL, R. R., DAVIES, J. H., MULVEY, J. H., and PERKINS, D. H., 1952, *Phil. Mag.*, **43**, 753.
DUERDEN, T., and HYAMS, B. D., 1952, *Phil. Mag.*, **43**, 717.
FRIEDLANDER, M., HARRIS, G. G., and MENON, M. G. K., 1953, *Proc. Roy. Soc.*, *Report of Discussion on V-Particles and Heavy Mesons*, in the press.

VON FRIESEN, S., and KRISTIANSSON, K., 1950, *Nature, Lond.*, **166**, 686; 1952, *Ark. Fys.*, **4**, 505.
FRY, W. F., and LORD, J. J., 1952, *Phys. Rev.*, **87**, 533.
ISACHSEN, N., VANGEN, V., and SÖRENSON, S. O., 1953, *Phil. Mag.*, **44**, 224.
LAL, D., PAL, Y., PETERS, B., and SWAMI, M. S., 1952 a, *Phys. Rev.*, **87**, 545; 1952 b, *Proc. Ind. Acad. Sci.*, **36**, 75.
LEVI-SETTI, R., and TOMASINI, G., 1952, *Nuovo Cim.*, **9**, 1244.
O'CEALLAIGH, C., 1951, *Phil. Mag.*, **42**, 1032.
POWELL, C. F., 1952, Report of International Physics Conference, Copenhagen, p. 29; 1953, *Phil. Mag.*, **44**, 219.
SALVINI, G., and KIM, Y., 1951, *Phys. Rev.*, **85**, 921.
SALVINI, G., and REYNOLDS, G., 1951, *Phys. Rev.*, **83**, 198.
VOYVODIC, L., 1952, *Phys. Rev.*, **86**, 1046.

S-Particles

ANNIS, M., BRIDGE, H. S., COURANT, H., OLBERT, S., and ROSSI, B, 1952 c, *Nuovo Cim.*, **9**, 624.
ANNIS, M., BRIDGE, H. S., and OLBERT, S., 1953, *Phys. Rev.*, in the press.
BRIDGE, H. S., and ANNIS, M., 1951, *Phys, Rev.*, **82**, 445.
BRIDGE, H. S., PEYROU, C., ROSSI, B., and SAFFORD, R., 1953 b, *Phys. Rev.*, in the press.

Charged V-Particles

ARMENTEROS, R., BARKER, K. H., BUTLER, C. C., CACHON, A., and CHAPMAN, A. H., 1951, *Nature, Lond.*, **167**, 501.
ARMENTEROS, R., BARKER, K. H., BUTLER, C. C., CACHON, A., and YORK, C. M., 1952, *Phil. Mag.*, **43**, 597.
ASTBURY, J. P., BUCHANAN, J. S., CHIPPINDALE, P., MILLAR, D. D., NEWTH, J. A., PAGE, D. I., RYTZ, A., and SAHIAR, A. B., 1953, *Phil. Mag.*, **44**, 242.
ASTBURY, J. P., CHIPPINDALE, P., MILLAR, D. D., NEWTH, J. A., PAGE, D. I., RYTZ, A., and SAHIAR, A. B., 1952, *Phil. Mag.*, **43**, 1283.
BARKER, K. H., BUTLER, C. C., SOWERBY, M. G., and YORK, C. M., 1952, *Phil. Mag.*, **43**, 1201.
BARTLETT, M. S., 1953, *Phil. Mag.*, **44**, 249.
BRIDGE, H. S., and ANNIS, M., 1951, *Phys. Rev.*, **82**, 445.
DANIEL, R. R., DAVIES, J. H., MULVEY, J. H., and PERKINS, D. H., 1952, Report by C. F. Powell at International Physics Conference, Copenhagen, p. 29.
EVANS, G. R., GRIFFITHS, T. C., EVANS, W. H., and MULLER, T., 1951, Report of Conference on V-Particles and Heavy Mesons, Bristol, p. 33.
FRETTER, W. B., 1951, *Phys. Rev.*, **83**, 1053.
HOLE, N., 1947, *Ark. Mat. Astr. Fys.*, **3B**, No. 12.
KOFOED-HANSEN, O., 1951, *Phil. Mag.*, **42**, 1411.
LEIGHTON, R. B., and WANLASS, S. D., 1952, *Phys. Rev.*, **86**, 426.
LEIGHTON, R. B., WANLASS, S. D., and ALFORD, W. L., 1951, *Phys. Rev.*, **83**, 843.
PAIS, A., 1952, Report on International Physics Conference, Copenhagen, p. 34.
PEIERLS, R., 1935, *Proc. Roy. Soc.* A, **149**, 467.
ROCHESTER, G. D., and BUTLER, C. C., 1947, *Nature, Lond.*, **160**, 855.
SERIFF, A. J., LEIGHTON, R. B., HSIAO, C., COWAN, E. W., and ANDERSON, C. D., 1950, *Phys. Rev.*, **78**, 290.
THOMPSON, R. W., COHN, H. O., and FLUM, R. S., 1951, *Phys. Rev.*, **83**, 175.
WILSON, J. G., and BUTLER, C. C., 1952, *Phil. Mag.*, **43**, 993.